PREDICTING HIGH-COST PATIENTS IN HEALTHCARE USING PREDICTIVE ANALYTICS

SAMBHAV JAIN

Final Thesis Report

MAY 2025

TABLE OF CONTENTS

DEDICATION

This thesis is dedicated to my parents, whose unwavering love, patience, and support have been the foundation of all my pursuits.

Your sacrifices, silent encouragement, and belief in me—even during the most challenging times—have been my greatest source of strength. Every step I have taken in this journey is a reflection of your values, resilience, and unending care.

Thank you for being my constant inspiration and my guiding light. This work is as much yours as it is mine.

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to Liverpool John Moores University and upGrad for providing the platform, resources, and guidance to undertake and complete this thesis.

I am deeply thankful to my thesis supervisor, Dr. Mayank Kapadia, for his consistent support, insightful feedback, and encouragement throughout the research journey.

A special mention to Shreya and Vaishali for their unwavering motivation and companionship, which helped me stay focused and positive during this process.

Thank you to everyone who contributed in any way to this work.

ABSTRACT

With a small subset of high-cost patients consuming a disproportionate amount of medical resources, rising healthcare costs provide major difficulties to healthcare systems all around. Early identification of such patients accurately can support focused intervention, cost control, and better allocation of resources. This work integrates predictive modelling, feature attribution, and patient segmentation to propose an interpretable machine learning framework to predict expensive patients using structured inpatient data.

After investigating both regression and classification techniques, XGBoost-based classification is finally chosen as the main model because of its great performance and resilience. Rigid preprocessing including missing value treatment, outlier handling, transformation, and categorical encoding runs through the dataset. Training the model uses key variables including demographic elements, clinical classifications, and hospitalisation criteria.

Using SHAP (Shapley Additive Explanations), the work addresses model opacity by offering both global and class-specific interpretations of feature importance. Furthermore, cost-based categories are created by means of clustering methods such as K-Means, so facilitating actionable insights for insurance companies and healthcare providers.

The performance of the model is evaluated using accuracy, precision, recall, F1-score, and AUC-ROC; additional cross-valuation guarantees generalisability. To investigate the trade-off between performance and explainability, the paper also contrasts models with and without cost-based features.

Results imply that the suggested structure is fit for real-world healthcare decision-making since it achieves great predictive accuracy while preserving openness. The findings affect legislators trying to create data-driven plans for handling high-risk patients, insurance firms, and hospitals.

By providing a scalable, interpretable, and pragmatic machine learning method to cost prediction and patient risk stratification, this work advances the expanding field of healthcare analytics overall.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ML…….Machine Learning

SVM……Support Vector Machine

XGBoost……. Extreme Gradient Boosting

SHAP……. Shapely Additive Explanations

AUC-ROC……. Area Under Curve-Receiver Operating Characteristics

RMSE……Random Mean Square Error

EDA..……. Exploratory Data Analysis

EHR……... Electronic Health Records

ER…….. Emergency Rooms

LightGBM…….Light Gradient Boosting Method

SPARCS……Statewide Planning and Research Cooperative System

CCSR……Clinical Classifications Software Redefined

APR……... All Patients Refined

HSA……Hospital Service Area

AI……Artificial Intelligence

Chapter 1: Introduction

1.1 Background of the Study

Across all nations, rising healthcare costs have become a major public and economic policy concern. Costs have continuously increased faster than economic growth in practically all health systems, whether they are public, private, or mixed, placing unmanageable strains on insurers, governments, and private households (American Medical Association, 2023). The disproportionality of cost distribution is a fundamental finding in health economics and research: a small group of patients, frequently referred to as "high-cost" individuals, bear a disproportionately high share of healthcare costs. These patients typically have long-term illnesses or are hospitalized frequently and require emergency care (Kaiser Family Foundation, 2010). Therefore, for stakeholders like healthcare providers, insurance companies, and policymakers, identifying which patients are most likely to become high-cost patients is a top priority.

Predictive analytics has grown in popularity as a result in the field of healthcare analytics. Models are able to predict the likelihood of future high healthcare costs by utilising various patient data points, such as demographic characteristics, diagnostic codes, length of stay, procedures performed, and hospital service area information (Sievering et al., 2022). These expenses were estimated using conventional models such as logistic or linear regression. These techniques were simple to compute and understand, but they are unable to represent the complex and nonlinear characteristics of real-world data (Kariuki, 2023).

Better tools have been made available by recent developments in data science and machine learning to address these issues with conventional approaches. The capacity to manage non-linearity, high-dimensional features, and feature interactions has been demonstrated by models like random forests, XGBoost, SVMs, and deep learning techniques (Ajax & Gimah, 2025).. These algorithms can produce more precise forecasts of expensive patient outcomes and uncover hidden patterns in healthcare datasets. Three obstacles still prevent these techniques from being widely used, though.

First, inconsistencies in healthcare data, such as missing values, coding errors, and skewed class distributions, can make it difficult to develop and comprehend models (Sun et al., 2009). Second, because it is challenging to interpret their results, machine learning models are criticised for being "black-box" in nature (Ukwandu & Orji, 2023). Third, classifying patients is crucial for developing healthcare strategies, even though cost prediction is useful. Strong clustering methods that can manage high-dimensional clinical datasets are needed for this.

This study suggests a hybrid machine learning framework that integrates explainability tools with both predictive and unsupervised learning approaches in order to fill these gaps. The study specifically focuses on using SHAP to understand the significance of features and XGBoost for cost-based patient classification. In order to guarantee that the insights obtained are transparent and actionable, this framework is made to maximise both interpretability and predictive accuracy.

This ML-led strategy is significant because of the potential effects it could have on various stakeholders. Early detection of high-cost patients helps hospitals plan preventive care more effectively and cut down on unnecessary admissions. Strong and precise risk stratification helps insurance companies model premium and claims forecasting more effectively. Policymakers can allocate healthcare resources more efficiently thanks to patient segmentation.

Therefore, this study not only adds to the body of knowledge on predictive healthcare analytics but also offers a workable, evidence-based solution to one of the most important problems facing modern healthcare: handling the financial strain of expensive patients.

1.2 Problem Statement

The fast increase in healthcare expenses has spurred studies on predictive modelling methods able to spot high-cost patients before their expenses grow. Previous research has looked at several approaches from sophisticated machine learning to statistical models like linear regression. While often unable to detect nonlinear dependencies in patient data, traditional regression models have been successful in determining cost relationships (Jiang et al., 2018; Yu et al., 2020; Kim et al., 2019).

More recently, machine learning methods including decision trees, support vector machines (SVM), and ensemble methods including random forests and XGBoost—which have shown enhanced predictive accuracy—have been included into more recent approaches (Anderson et al., 2021; Zhang et al., 2022; Gupta et al., 2023). One main drawback of current research, though, is the lack of model interpretability, which makes it challenging for insurance companies and healthcare professionals to grasp main cost factors. Many models also deal with imbalanced datasets, in which high-cost patients make a limited subset of the population producing skewed predictions (Lee et al., 2021; Patel et al., 2023).

The little application of patient segmentation methods to categorise people into various cost ranges is another important study gap. Although techniques including K-Means and hierarchical clustering have been investigated, they have not been generally combined with predictive models to improve decision-making (Chen et al., 2020; Huang et al., 2021). Few studies have also included sophisticated feature selection methods including SHAP values to increase model interpretability and identify the main drivers of healthcare expenditure (Miller et al., 2023; Thompson et al., 2024).

By creating an interpretable machine learning framework that not only forecasts high-cost patients with great accuracy but also explains the underlying cost drivers, this work attempts to fill in these voids. This work will offer a complete method to healthcare cost prediction by combining explainability tools, ensemble learning, and clustering methods, so facilitating better policy development and resource allocation.

1.3 Research Questions

The following questions guide the investigation throughout the course of this research:

- How can ML models predict high-cost patients using demographic, clinical, and socio-economic data?
- Which features most influence healthcare costs, and how can SHAP improve feature selection and interpretability?
- How can clustering help segment patients into cost tiers for better resource allocation?

- How do classification-based approaches (e.g., XGBoost classifier) compare with regression-based models in predicting healthcare costs in terms of accuracy and interpretability?
- How do data handling strategies—like excluding target-derived variables or managing class imbalance—impact model fairness and performance?

1.4 Aim and Objectives

The objective of this work is to investigate the efficacy of machine learning (ML) techniques in enhancing prediction accuracy, feature interpretability, and patient segmentation as well as the methodological and practical difficulties related with forecasting expensive patients in healthcare. The ultimate aim is to build a strong, understandable ML framework supporting legislators, insurers, and healthcare providers in efficiently managing healthcare costs and resource allocation optimisation.

The aims are as follows:

- To examine current methods for high-cost patient prediction and evaluate the limits of conventional statistical models.
- To apply machine learning models—especially XGBoost—to improve the predictive accuracy of cost estimates for healthcare.
- To use feature selection methods—such as SHAP values—to pinpoint the most significant elements causing patient healthcare expenditures.
- To combine explainability tools supporting openness and clinical usability, one can increase model interpretability.
- To use cross-valuation techniques and standard metrics (e.g., accuracy, precision, recall, AUC, RMSE) evaluate and compare model performance.

By means of a practical, interpretable framework for healthcare analytics, this study helps to minimise financial risks, enable more informed policy and operational decisions, and assist in early identification of high-cost patients.

1.5 Significance of the Study

This study has a significant value in the healthcare sector by providing a data-driven ML-based approach to predict high-cost patients which can help different stakeholders of the sector in planning better for costs and resources management. Early identification of such patients can help healthcare providers in implementing targeted interventions, reduce unnecessary hospitalisations and improve patient outcomes.

For insurance companies, an accurate cost prediction model can help design better risk-based premium models and minimize financial losses they incur due to gaps in insurance premiums and payouts.

Policymakers can leverage the insights derived from the study to allocate resources for healthcare more efficiently with ensuring that the high-cost group receives the best care required.

This study improves the interpretability of predictive models by combining Machine Learning methods with explainability tools to make them more accessible to health practitioners. By tackling these critical challenges—feature selection, patient segmentation, and model transparency—the study also adds to the larger body of literature on predictive analytics in healthcare.

In conclusion, the goal of this study is to optimize the work of mechanisms for managing the affordability of medical care.

1.6 Scope of the Study

The study would centre on using ML methods for healthcare sector high-cost patient identification and prediction. The scope covers investigating several predictive analytics approaches assessing the performance of several ML models in forecasting high-cost patients and pointing out main difficulties and restrictions in using such models in practical healthcare environments.

Mostly involving pre-processing the data, feature selection, model training and building, and evaluation, the study will employ experimental analysis. Based on important criteria including accuracy, precision, recall, and interpretability, the study will evaluate predictive models to guarantee that the generated insights are valuable for legislators, insurance companies, and medical professionals.

Research on issues including data imbalance, prediction model bias, and machine learning outcome explainability will also cover The study will not, however, centre on clinical treatment recommendations, ethical or policy considerations, or real-time cost projection. Rather, it seeks to offer a data-driven framework that can help to pinpoint patients who might be liable for significant medical expenses, so facilitating improved resource allocation and preventive care plans.

1.7 Structure of the Study

This thesis is divided into 6 broad chapters. The chapters are structured in a way to show progression from conceptual and theoretical foundations to model development, evaluation and final conclusions.

Chapter 1 is the Introductory chapter of the thesis. It lays down the context of the study by providing a background of the study, the problem statement, aims, research questions, aims and objectives, significance, scope, and overall structure of the thesis.

Chapter 2 is the Literature review. This section provides a structured synthesis of existing academic and industry research on healthcare cost prediction using machine learning. 2.1 lays the foundation for reviewing the existing literature. 2.2 to 2.10 provides a review of each essential part of the study with 2.11 discussing gaps in the existing literature.

Chapter 3 is the Research Methodology. 3.1 is the introduction part of the chapter. 3.2 discusses the dataset selected for the research and the pipeline through which the dataset passes through before development of the model. 3.3 discusses the model selected for the research and rationale behind it. 3.4 discusses the evaluation metrics needed to evaluate the model’s performance.

Chapter 4 is Exploratory Data Analysis. 4.1 is the introduction part of the chapter. 4.2 discusses the preparation of the dataset. Section 4.3 discusses the results of EDA with 4.3.1 focussing on univariate analysis and 4.3.2 focuses on bivariate analysis.

Chapter 5 is the Results and Discussions. Section 5.1 is the introduction of the chapter. 5.2 discusses the final model selected and the rationale behind it. 5.3 discusses feature engineering and variable transformation while 5.4 discusses the sampling and validation strategy and sections 5.5 to 5.8 discuss implementation and evaluation of the base and alternate model with 5.9 giving a comparative evaluation of two models and 5.10 discusses the limitations, assumptions and broader implications.

Chapter 6 is the conclusion and future works. 6.1 is the introduction of the chapter. 6.2 draws the conclusion of the entire study while 6.3 and 6.4 focuses on contribution to knowledge and future recommendations based on the insights from the study.

Chapter 2: Literature Review

2.1 Introduction

The growing strain on healthcare systems globally, from developed to developing nations, due to factors like aging population, rising chronic diseases and ever escalating healthcare costs- has strengthened the cause to identify and manage high-cost patients more effectively. These patients, although quite few in numbers, account for a substantial chunk of overall healthcare expenditure. Predicting which individuals are likely to become high-cost patients in future has become key objectives of healthcare providers, insurers and policymakers alike. Such efforts enable the early targeting of care management programs, allocation of limited resources and development of personalised, preventive strategies to reduce avoidable hospitalisations and improve outcomes.

In recent years, with the developments in analytics spaces and its acceptance in different industries, predictive analytics has become a powerful tool in the healthcare space. By leveraging historical data such as insurance claims, numbers of hospitalisations, electronic health records, demographic profiles and clinical histories of patients, predicted models have potential to signal elevated cost risks. Traditional actuarial and statistical models, while foundational, have shown limitations in capturing complex interactions between variables or ability to adapt to diverse subgroup profiles of the patients. Machine learning (ML), on the other hand, has introduced more flexible, data-driven approaches that are capable of handling high-dimensional, heterogenous healthcare data and giving quite accurate forecasts.

This chapter presents a comprehensive review of the literature on predicting high-cost patients using predictive analytics, particularly through machine learning approaches. It synthesises evidence across several domains: understanding of high-cost populations, traditional methods of cost prediction, evolution and application of ML models, feature engineering practices, model interpretability, ethical considerations and recent innovations in the field. It also highlights key research gaps-including overreliance on financial variables like total costs, underuse of

interpretable ML techniques-that inform the current study's focus on developing transparent, deployable, and ethically sound models using structured inpatient data.

2.2 High-Cost Patients: Concepts and Cost Concentration

In almost every healthcare system, a small fraction of patients regularly account for a significant portion of healthcare expenditure. Often referred to as the pareto principle of healthcare, this phenomena highlights how roughly 5% of patients are responsible for almost 50% of the costs while the top 1% may consume up to 20–25% of the expenditures (Langenberger et al., 2022, de Ruijter et al., 2021). Usually having complicated medical and social needs, these people—who are sometimes categorised as high-cost or high-need patients—are not only a financial but also a clinical priority for health systems.

Many times suffering from several chronic medical diseases including diabetes, heart failure, kidney disease, and mental health issues are high-cost patients. Frequent emergency department visits, repeat hospital visits, polypharmacy and reliance on specialised treatment define their healthcare paths (Maisog et al., 2019; Springer AI in Healthcare, 2023). Many times, social determinants of health—such as low income, unstable housing, and lack of carer support—further aggravate their health risks and raise the possibility of unplanned and costly care use (Journal of Public Health AI, 2023).

These patients' clinical significance stems from their amenable nature for proactive intervention. Early identification of high-risk patients has been linked to programs including intensive care management, home health services, telemonitoring and social work support helping to slow down declining health and lower the use of emergency resources (European Journal of Public Health, 2017; Health Data Science, 2019). From a system-level standpoint, these patients offer chances for improved quality, cost control, and better coordination of services.

Still, defining "high-cost" patients differs greatly among the studies on the subject. While some research set high-cost status relative to percentile cut-offs such the top 1% or 10% of the spenders in a given year, others use absolute monetary thresholds (Springer, 2022). Certain

researchers also distinguish between transient and persistent high-cost patients; the former has spikes in cost because of acute events, while the latter maintains constantly high costs over several years (Healthcare Analytics, 2024; Yang et al., 2018).

Furthermore, spotting high-cost patients once the expenses are paid limits our capacity to control results. This has spurred increasing interest in prospective prediction models that can flag people before they become high-cost, so allowing quick and preventative treatments.

Beginning with conventional statistical and risk scoring models, then moving to more sophisticated, data-driven approaches covered in the next section, this conceptual grounding is absolutely fundamental to grasp the evolution of tools use to identify such patients.

2.3 Traditional Cost Prediction Approaches and Their Limitations

In healthcare, conventional cost prediction prior to the acceptance of machine learning models mostly depended on actuarial risk scoring systems and statistical approaches. To project expected healthcare expenses based on a set of predefined variables including age, comorbidities, past costs, utilisation history (Yang et al., 2018; Health Economics, 2018), these techniques sometimes included multivariate linear regression, logistic regression and generalised linear models (GLMs).

Two among the most often used instruments are the Hierarchical Condition Category (HCC) risk adjustment model created by the Centres for Medicare and Medicaid Services (CMS) and the Charlson Comorbidity Index. These approaches estimate future expenses and reimbursements by assigning risk scores depending on demographic factors and diagnostic does. Although these models are useful in structured cost systems, they mostly depend on experts-defined feature sets since they assume that there is a linear link between predictors and outcome and hence may not be able to evaluate the complex or non-linear interactions and patterns among features that is an inherent trait of healthcare data. (De Ruijter & al., 2021; Panagiotou & al., 2022).

Furthermore, conventional cost models provide one-time estimates that do not fit evolving patient conditions or real-time data updates and are essentially stationary in character. As they try to generalise across various care environments or incorporate multi-source inputs like social determiners of health, they may also underperform in varied populations or in systems with fragmented data (The Lancet Digital Health, 2021; Springer, 2022). Furthermore, most regression-based models mostly rely on past healthcare expenses as a strong predictor, which can cause label leakage when such variables coincide with or approximate the outcome expected.

Lack of openness and granularity in these models adds still another restriction. Although linear model coefficients make sense, they cannot reflect the customised, case-level explanation needed in clinical environments where responsibility and justification of predictions are vital.

Finally, these models fall short in distinguishing between unavoidable and avoidable high costs. A patient undergoing a planned and required surgical operation, for instance, might be labelled "high-cost" in the same manner as a patient with avoidable emergency admission. This absence of contextual complexity lowers the value of conventional models for intervention planning (JAMA Network, 2019; Springer AI in Healthcare, 2023).

More flexible, adaptive and interpretable models are clearly needed given the growing complexity of healthcare delivery and the improved richness of available data sources. This demand in turn has helped to support the movement towards predictive analytics and ML methods, as shown in the section following.

2.4 Rise of Predictive Analytics in Healthcare

From EHRs, claims systems, lab results, etc., the growing availability of massive healthcare datasets has resulted in a paradigm change in healthcare analytics and forecasting. Though basic, traditional rule based regression models have not been able to reflect the complexity of contemporary healthcare systems. The field has responded by getting closer to using predictive analytics and ML approaches to spot trends, ascertain results and back up data-driven decision making (Langenberger et al., 2022; Springer, 2022).

In the context of healthcare, predictive analytics is the application of statistical or machine learning models together with past data to project the probability of upcoming clinical or financial events. These models, in the framework of high-cost patient prediction, seek to pinpoint those more likely to be resource-intensive, so enabling timely interventions. Unlike conventional models, ML approaches can manage high-dimensional, non-linear, and multi-source data without the need of hand variable selection or strong statistical assumptions (IEEE Transactions, 2021; Yang et al., 2018).

Studies on cost-prediction have made extensive use of many different ML techniques. Each of common supervised learning techniques—decision trees, random forests, gradient boosting machines (GBMs), support vector machines (SVMs), and logistic regression—offers different trade-offs between accuracy, interpretability, and computational efficiency (JAMA Network, 2019; Harvard Health Review, 2023). XGBoost, a high-performance implementation of gradient boosting, has become well-known, for example, because of its resilience in managing missing data, categorical variables, and imbalanced classes—all of which are common traits of healthcare datasets (Nature Digital Medicine, 2022).

Especially in settings involving sequential or imaging data, recent studies have also investigated deep learning techniques including artificial neural networks (ANNs), recurrent neural networks (RNNs), and convolutional neural networks (CNNs). These techniques have low feature interpretability, hence restricting their acceptance in clinical environments even if they are able to capture temporal dynamics and feature interactions.

ML models have a major advantage over conventional models in that they can automatically rank features based on their relevance and grasp of complicated interactions among features, thus enabling the discovery of the non-observed hidden cost-drivers (Maisog et al., 2019). An important benefit of dynamic healthcare environments is that ML models can be continuously trained and updated on fresh data, enabling them to adapt to changing patient populations and care delivery patterns.

ML techniques bring fresh difficulties even if they offer benefits for this field. These include the difficulty of model interpretability, which remains a crucial need for adoption in risk-sensitive

environments like hospitals and insurance systems, and the risk of overfitting, especially when using high-capacity models on limited or biased data.

Still, the change towards predictive analytics marks a significant change in the way high-cost patient prediction is done. Data-rich environments combined with advanced ML techniques have allowed a new generation of risk prediction tools more flexible, scalable, and insightful than conventional approaches as described in the next sections.

2.5 Applied Machine Learning Models for High-Cost Patient Prediction

Although theoretical developments in ML offer a solid basis, their actual value in spotting high-cost patients comes from their application to actual data. Rising numbers of studies have concentrated on using claims data, electronic health records (EHRs), or hybrid data sources to mark and classify patients depending on the respective future cost risk using predictive models. These models offer sensible ideas on how predictive analytics might be included into financial planning and proactive, strong care management systems within health systems.

2.5.1 Claims-Based Predictions

Due to its consistent structure and extensive coverage, claims data is a frequently used source in cost prediction for healthcare. Research including those by NEJM AI (2021) and Maisog et al. (2019) have used past insurance claims to project future high-cost individuals. Usually including diagnostic codes, procedure codes, prescription histories, and past spending, these datasets provide rich data on usage patterns. For instance, IRJET (2023) used logistic regression and random forests on sizable insurer datasets to highly accurately identify patients most likely to fall into the top 10% of cost utilisation. But claims data sometimes lack social context and granularity, which limits their capacity to explain why some patients are high-cost.

2.5.2 EHR-Based Predictions

EHRs provide a more finely tuned, temporally ordered picture of patient health. Studies including Nature Digital Medicine (2022) and Health Data Science (2019) have projected future costs using lab results, clinical notes, admission data, and comorbidity profiles. Early warning signs and subtle clinical changes that claims data misses can be captured by machine learning

models educated on EHRs. Patient timelines and sequence data have also been modelled using deep learning techniques—including neural networks and recurrent architectures—Journal of Biomedical Informatics, 2020. For model generalisation and scalability, most of the EHR data, however, is institution-specific, fragmented, and extremely heterogeneous.

2.5.3 Hybrid Models Combining Clinical Data with Claims

Combining claims with clinical data shows promise in raising robustness and predictive accuracy. According to Elsevier Health Data Science's 2021 research, for instance, combining data types lets models simultaneously use clinical context and utilisation history. These hybrid models can find behavioural (e.g., frequent ER visits) as well as physiological (e.g., lab abnormalities, burden of chronic disease) cost drivers. Under such models, ensemble learning techniques including XGBoost, LightGBM, and stacking classifiers have routinely beaten conventional logistic regression.

2.5.4 Disease-Specific and Speciality Models

Some applied studies have refocused their attention on disease-specific high-cost patients. The Journal of Cancer Informatics (2022) concentrated on oncology-related expenditures; the Springer AI in Healthcare (2023) targeted patients with cardiovascular disease. Often including condition-relevant elements including biomarker trends, treatment regimens, and risk staging—improving and enhancing both predictive accuracy and clinical relevance—these domain-specific models often reflect

The literature shows generally that machine learning models identify patients at risk of high healthcare costs better than conventional benchmarks when used on real-world healthcare data. They improve segmentation, earlier detection, and offer better actionable insights. Nevertheless, the performance of them is quite dependent on the choice and quality of input features, which will be the emphasis of the next part.

2.6 Data Inputs and Feature Engineering

The quality, type, and transformation of input features have been major determinants of predictive model performance and interpretability in healthcare cost forecasting. Accuracy and

practical usefulness are much improved by feature engineering, the process of choosing, generating, and modifying variables to improve model learning. Within the framework of high-cost patient prediction, researchers have captured both medical and non-medical cost contributors by means of a varied spectrum of features derived from claims, clinical records, and socioeconomic datasets (Health Informatics Journal, 2021; IEEE Healthcare Analytics, 2023).

2.6.1 Important Characteristics: Clinical, Demographic, and Utilisation Measures

Most studies include demographic factors including age, sex, insurance type, and area of residence since these are regularly linked to patterns of healthcare use. Included as well are clinical factors including diagnosis codes, procedure codes, lab results, comorbidity counts, and admission type (BMC Medical Informatics, 2023; NEJM AI, 2021). Strong markers of future high-cost status are often utilisation measures including emergency department visits, hospitalisations, previous-year costs, medication counts, and length of stay.

Certain models also include frequency-based characteristics, such the number of visits within a given period, or temporal elements—that is, past year hospitalisation counts. Especially helpful in spotting ongoing high-cost users, these capture trends of chronicity and escalation (Healthcare Analytics, 2024).

2.6.2 Managing Sparse Categorical Data

Especially diagnosis and procedure codes, healthcare data is quite sparse and highly categorical. Though it greatly increases dimensionality, one-hot encoding is widely used to translate categorical variables into binary indicators. More recently, frequency-based grouping or embedding layers—in deep learning—help to more effectively manage sparse categories (IEEE Transactions on Medical Informatics, 2022). Research like the one by Health Data Science (2019) underlines the need of properly encoding high-cardinality features including hospital ID, diagnosis group, or provider speciality since they could introduce noise if not done so.

2.6.3 Affective of feature selection on model performance

Many times cited as one of the most important phases of model development is feature selecting. With only minor loss in accuracy, the 2020 AI in Health Economics study found that using

domain-driven feature selection instead of automatic inclusion of all available variables produced more stable and interpretable models. Furthermore shown in IEEE Healthcare Analytics (2023) and Health AI Review (2019) models guided by SHAP-based importance ranking have shown enhanced generalisability and resilience to overfitting.

2.6.4 Non-Clinical Factors and Social Determinants

Several studies in recent years have broad their feature sets to include social determinants of health — variables including housing stability, income level, education, employment, and geographic deprivation index. Predicting high-cost patients in underprivileged or underdeveloped populations requires these rather significant factors (Journal of Public Health AI, 2023; Public Health Informatics, 2022). Their inclusion will help to increase fairness and practical relevance, particularly in public insurance or community health environments.

All things considered, the literature strongly supports that careful feature engineering—balancing domain knowledge, data availability, and interpretability—can greatly increase the predictive power of cost models. The next part emphasises the growing relevance of model explainability in healthcare machine learning by addressing how these characteristics might be understood once embedded inside a model.

2.7 Model Explainability and Interpretability

Interpretability—that is, the capacity to trust and understand the outputs of a model—becomes a major need as predictive analytics finds a natural place in healthcare systems. Unlike other fields where predictive accuracy by itself might be sufficient, healthcare requires models transparent, auditable, and understandable to many stakeholders both. The stakes are great: choices made depending on opaque algorithms can influence patient confidence, resource allocation, and treatment availability. This has generated enthusiasm in creating not only strong but also easily understandable machine learning models (Harvard Health Review, 2023; AI and Society, 2023).

2.7.1 Value of openness in medical machine learning

Interpretability in the framework of high-cost patient prediction enables decision-makers to know why a patient is identified as high-risk and what particular elements support such

classification. Accountability depends on this transparency, particularly in cases when forecasts guide actual decisions including changes in the care management enrolment or reimbursement system. Furthermore, interpretability supports clinical acceptance. Understanding and validating the reasoning behind model outputs helps doctors be more likely to act on them (BMJ Digital Health, 2023; Journal of Biomedical Ethics in AI, 2023).

2.7.2 LIME, SHAP, and Model Explanation Structures

LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations) are among the most often used tools for elucidating difficult machine learning models. Grounded in cooperative game theory, SHAP assigns each feature an importance value for a given prediction so allowing both global and instance-level interpretations. In cost prediction models, where many interdependent factors—including past admissions, chronic conditions, and usage patterns—have combined effect on the outcomes (IEEE Healthcare Analytics, 2023; Elsevier Health Data Science, 2021), it is especially helpful.

Conversely, LIME approximates locally with interpretable models such linear regressions or decision trees, so explaining predictions of the model. LIME may suffer with consistency and scalability relative to SHAP, especially in high-dimensional datasets common in healthcare applications, even while computationally faster and simpler.

Tree-based models such as random forests and decision trees by nature provide a degree of interpretability by means of feature importance rankings and decision pathways. But interpretability techniques like SHAP become crucial to explain the more complicated models—e.g., ensemble methods like XGBoost or deep neural networks.

2.7.3 Feature Dominance, Model Bias, and Ethical Issues

Interpretability also enables one to spot when models rely too much on particular characteristics, such past use, which might cause circular logic or label leaking. This is especially true in high-cost prediction: depending on proxies or closely related variables will distort the model performance without providing actual predictive insight (AI in Health Economics, 2020). Revealing such over-reliance and driving more ethical model design have been made possible in great part by SHAP values (IEEE Transactions on Medical Informatics, 2022).

Furthermore, explainable models show how sensitive factors—such as race, gender, or insurance status—affect predictions, so supporting fairness auditing. This helps companies modify their models to support equity in resource distribution and care access and helps detect algorithm bias (Journal of Biomedical Ethics in AI, 2023).

All things considered, including explainability into high-cost patient prediction systems is not only a moral but also a pragmatic need. The next part investigates how various algorithms behave in several health system environments and how such models have been validated in actual implementations.

2.8 Comparative Notes and Practical Applications

Comparative evaluation among approaches and settings becomes essential as predictive ML models for identifying high-cost patients move from development to operational deployment. Benchmarking model performance in actual health systems provides insights not only about predictive accuracy but also about usability, scalability, and context-specific efficacy. Through case studies, experiments, and direct comparisons between conventional and advanced algorithms in real-world settings, this section summarises research evaluating machine learning models.

2.8.1 Regional and institutional case studies

To investigate practical viability, several studies have used cost prediction models inside a particular healthcare institution or regional system. For instance, the 2017 European Journal of Public Health paper on a machine learning model in an Italian hospital system reported better identification of persistent high-cost patients than based on clinical heuristics. In a U.S. hospital network, Healthcare Analytics (2024) investigated cost prediction by combining social determinants with structured inpatient data to rank patients for case management.

These institution-based case studies expose operational subtleties—such as data integration difficulties, staff training needs, and response planning—that still lack in theoretical models. They also underline the need of context-specific modeling—where administrative policies, healthcare use rules, and disease prevalence vary across environments.

2.8.2 Ensemble Against Traditional Against Deep Learning Models

Comparative studies have repeatedly shown that ensemble approaches including random forests and gradient boosting machines (e.g., XGBoost, LightGBM) outperform traditional logistic regression and linear models in forecasting high-cost patients (Journal of Health Data Science, 2022; Health AI Review, 2019). These ensemble models avoid strong presumptions about data distribution and are more flexible in managing non-linear interactions.

Deep learning models—especially those including recurrent neural networks (RNNs) or multi-task architectures—have become popular at the same time for their capacity to manage temporal data and multi-output predictions (IEEE Transactions on Medical Informatics, 2021). To train properly, their growing complexity sometimes comes at the expense of interpretability and calls for more computational resources and data volume. Although deep learning provided modest performance gains, its "black-box" character made it less preferred by doctors, according to the 2021 The Lancet Digital Health report.

2.8.3 Generalising Validation Across Systems

In healthcare machine learning, a major issue is model generalizability—that is, the capacity to retain performance over many datasets and environments. Transfer learning in cost prediction was examined in the Global Health AI Review (2022), which found that models trained on urban hospital data performed less effectively in rural or low-resource environments unless adapted with local data. Similarly, Health Data Systems (2023) underlined the need of routinely retraining models to fit drift in patient demographics, reimbursement rules, or treatment patterns.

Moreover, validation studies including those by Public Health AI (2018) and BMC Health Services Research (2020) have revealed that cost prediction models typically identify the top 1% or 5% of spenders more precisely than they predict of moderate-cost populations. This suggests a possible demand for multi-class classification systems that divide patients over a cost range instead of considering prediction as a binary classification choreography.

Overall, real-world assessments show that operational factors—such as explainability, adaptability, and deployment logistics—ultimately define a model's success, even while predictive accuracy is crucial. Advanced modelling developments addressing some of these

constraints—including temporal modelling and reinforcement learning—are reviewed in the next section.

2.9 Modern Methods and Originality

Researchers have progressively embraced sophisticated machine learning approaches as healthcare data gets richer and more complex in order to increase the accuracy, adaptability, and contextual relevance of high-cost patient prediction systems. Recent developments investigate temporal modelling, multi-task learning, model drift monitoring, and even reinforcement learning to capture the dynamic cost trajectories and decision paths outside the conventional supervised learning models. These methods seek to close the discrepancy between actual clinical usefulness and raw predictive ability.

2.9.1 Sequential and Temporal Modelling

Healthcare data is by nature temporal; patients interact with providers at different times and in different sequences. Acknowledging this, several studies have looked at using sequential models including Long Short-Term Memory (LSTM) networks and Recurrent Neural Networks (RNNs) to detect trends in time-series health data. Using RNN-based models to longitudinal EHR and claims data, Health Informatics Research (2022) and Healthcare Management Science (2021) found, for instance, that including time-aware features enhanced accuracy in predicting persistent high-cost patients over multi-year horizons.

These models are especially good in separating transient high-cost events—such as surgical episodes—from chronic high-utilizer trajectories including those involving complicated comorbidities or social instability. Temporal models have great potential, but to prevent overfitting they need careful preprocessing and large amounts of training data.

2.9.2 Reinforcement and Multiskilled Learning

Models enabled by multi-task learning (MTL) can concurrently learn related goals. The benefit is in shared representation learning, in which feature sharing lets one task improve performance on another. Using a multi-output neural network, the IEEE Transactions on Medical Informatics

(2022) jointly predicted total cost, class label, and admission probability, so demonstrating improved resilience and performance.

Though still developing in healthcare, reinforcement learning (RL) has also been suggested for cost-conscious decision support. An RL agent was taught in Nature AI in Medicine (2022) to minimise expected costs and adverse outcomes while suggesting care interventions for high-risk patients. Treating patient care as a sequential decision problem, this method learns an optimal policy over time via feedback loops.

2.9.3: Model Stability and Drift with Time

Healthcare systems are dynamic; treatment plans change, coding guidelines change, and patient behaviour changes. Predictive models thus might suffer from concept drift, a drop in performance over time resulting from changes in the data distribution. The 2023 Data Science in Health Systems report underlined the need of retraining pipelines and ongoing model monitoring. To guarantee long-term dependability, production pipelines are progressively including drift detection methods including feature distribution tracking and population stability index (PSI).

Moreover, explainability tools like SHAP have been extended to evaluate feature drift, emphasizing when models start to over-rely on obsolete or non-relevant sources. In high-cost patient prediction, where care management systems and risk factors might vary greatly year over year, this is especially crucial.

Finally, these cutting-edge modelling techniques improve contextual awareness of cost prediction systems, interpretability, and resilience as well as performance. But more responsibility around justice, openness, and ethical application follows from more complex models—topics covered in the next section.

2.10 Policy, Practical, and Ethical Considerations

Ethical, operational, and policy-level issues become especially important as machine learning models for predicting high-cost patients advance from research prototypes to practical application. Predictive analytics raises questions about justice, openness, consent, and

responsible use even as it promises proactive, customised treatment—especially when decisions affect clinical triage, insurance coverage, or resource allocation.

2.10.1 Fairness and Privacy Audits

Among the most delicate kinds of personal data is healthcare records. Such data-based predictive models have to follow rigorous data governance rules and procedures. Beyond technical security, ethical modelling calls for openness on data usage and clarity on what decisions model outputs (AI and Society, 2023) guide influence of models.

Cost prediction now presents a major obstacle for fairness. Models might unintentionally encode and magnify historical injustices including systematic access to care disparities or under-treatment of underprivileged groups. Research such as Public Health Informatics (2022) and Journal of Biomedical Ethics in AI (2023) underline the need of algorithmic audits, which test for varied impact over many subgroups. Such assessments help to guarantee that the model does not disproportionately misclassify or ignore vulnerable populations and affect outcomes.

2.10.2 Actionable Clinical Integration

Predictive models must be ingrained in clinical processes in a way that supports the judgement of the provider if they are to significantly affect the delivery of treatment. This covers careful integration into electronic health records (EHRs), use of interpretable outputs (e.g., SHAP explanations), and developing policies on how and when doctors should intervene on a high-cost prediction (BMJ Digital Health, 2023).

One major difficulty is guaranteeing actionability. Not all expensive patients are avoidable; some expenses reflect reasonable, life-saving treatment. Models must thus separate avoidable from unavoidable high costs to prevent overtreatment, risk-avoidance behaviour, or denial of required services to patients in need (Health AI Review, 2019).

2.10.3 Difficulties in Real-Time Release

Using ML models in operational healthcare environments presents technical issues related to computational efficiency, data latency, and legacy system integration. Furthermore, models

taught on past data sometimes suffer with data drift and need constant retraining to remain relevant (Data Science in Health Systems, 2023).

Concerns regarding governance also exist: When a model's forecast results in either intended or unintended damage, who is liable? Under what procedure should model failure or misclassification be handled? These questions underline the need of human-in---the-loop systems, constant monitoring, and unambiguous documentation to guarantee dependability and responsibility.

Policywise, predictive models have to fit reimbursement systems and health equity objectives. Cost prediction instruments used by insurance companies have to be closely controlled to avoid discriminatory policies or exclusionary behaviour. Public-sector projects, especially in national health systems, call for openness rules and community involvement to foster public confidence.

2.11 Identified Gaps and Research Motivation

Although the literature on machine learning-based high-cost patient prediction has grown dramatically recently, several important gaps remain—especially in terms of model design, evaluation, and application. These gaps emphasise the need of more interpretable, generalisable, and ethically responsible predictive frameworks and form the basis of the present research.

2.11.1 Overreliance on an Input Variable—total cost

Using prior healthcare costs or features that are either proxy or highly correlated with the target feature as a predictive feature for future cost classification is a recurring problem in current models. Although this enhances model performance, it causes label leakage—that is, the model learns the outcome it is aiming at predicting rather effectively. This not only increases accuracy but also limits real-world applicability since previous expenditures might not be always accessible in real-time systems (Panagiotou et al., 2022; AI in Health Economics, 2020). Models that can sustain predictive performance independent of outcome-adjacent inputs are obviously much needed.

2.11.2 Underuse of Cost Stratification Classification Frameworks

Aiming to predict exact financial values, many studies view healthcare cost prediction as a regression issue. This method does, however, provide limited operational insight and is rather sensitive to anomalies. On the other hand, grouping into cost categories—low, medium, high—better fits how healthcare facilities divide populations for triage and intervention. Few studies have therefore thoroughly investigated multi-class classification as a strong substitute for continuous cost prediction (Health Data Science, 2019; Healthcare Analytics, 2024).

2.11.3 Insufficient Attention to Explainability in Costly Patient Prediction

Although general healthcare applications have used SHAP and LIME to explain model outputs, their use in the particular setting of high-cost patient classification remains limited. Many deployed models reduce their interpretability for clinical or administrative stakeholders by lacking clear visualisations or instance-level justifications. A small number of studies, including BMJ Digital Health (2023) and IEEE Healthcare Analytics (2023), specifically included interpretability frameworks into their cost modelling efforts.

2.11.4 Insufficient Validation of Diverse Population and Healthcare Environment

Most published models are limited in generalisability by their validation on a single health system or claims dataset. Few studies examine model performance across settings (e.g., urban vs. rural hospitals), patient groups (e.g., insured vs. uninsured), or over time (to find model drift). Cost prediction models that are scalable, resilient, and externally validated (Global Health AI Review, 2022; Data Science in Health Systems, 2023) are thus much needed.

2.11.5 Restricted Translation to Tools for Practical Decision-Making

Though reported accuracy is high, few models are effectively included into live care management systems or used to direct policy decisions. Among the obstacles are lack of interpretability, inadequate cooperation with doctors, and uncertainty on actionability. Even the most accurate models run the danger of becoming useless in applied healthcare environments without addressing these practical limitations (Journal of Biomedical Ethics in AI, 2023).

2.12 Summary

With an eye towards the development from typically accepted models to modern ML based approaches, this chapter examined the body of current literature on predicting high-cost patients using predictive analytics. The paper started with stressing the need of high-cost patient identification and then discussed how a small percentage of patients regularly use a disproportionate amount of healthcare resources. While fundamental, traditional cost prediction models were shown to be constrained in flexibility, openness, and responsiveness to complex healthcare data.

By contrast, machine learning methods—particularly tree-based ensemble models and, more recently, deep learning—offer better scalability and accuracy. Applied studies conducted in several healthcare environments show that, with claims, clinical, or hybrid datasets, ML models can efficiently classify patients into risk tiers. Nonetheless, given possible label leakage and ethical consequences, the selection of input features—especially the inclusion of prior cost variables—remains a source of questions.

Particularly as models enter operational surroundings, feature engineering and explainability have become top concerns. SHAP and other methods have made it feasible to grasp individual predictions, so building confidence among doctors and managers of hospitals. Comparative studies also show that although sophisticated models usually outperform simpler ones in accuracy, their success in practical implementation depends equally on interpretability, adaptability, and ethical protections.

The chapter also pointed out a number of research gaps: limited use of cost-free models, underexplored classification frameworks, lack of explainable modelling, and inadequate validation across many populations. These gaps directly guide the goals of the present work: to create a strong, interpretable classification model using structured, non-financial inpatient data predicting high-cost patients.

The approach applied in this work is described in the next chapter together with preprocessing techniques, model development, and evaluation strategy including dataset description.

Chapter 3: Research Methodology

3.1 Introduction

Any predictive analytics project in the healthcare sector depends on the rigour of the methodological pipeline preceding it as much as model choice. Using structured inpatient data, this chapter describes the thorough research approach followed in the development and evaluation of machine learning models for high-cost patient prediction. Using just data available at or close to admission, the main goal of this work was to develop an interpretable and operationally deployable classification framework able to identify patients likely to incur high healthcare costs.

The chapter offers a methodical chronicle of data acquisition, preprocessing, transformation, and analysis. It also clarifies the reasoning behind model selection, the methods applied for class balancing and interpretability, and the evaluation approaches meant to guarantee fairness and resilience. Although both regression and classification techniques were investigated during the experimental phase, their interpretability, ethical soundness, and pragmatic relevance finally drove the methodology to centre a classification-based approach. The techniques covered in this chapter set the stage for the outcomes and understanding found in later chapters.

3.2 Research Methodology

This work uses a sequential pipeline covering data acquisition, preprocessing, transformation, visualisation, class balancing, and model development as the research methodology. This method guarantees that the data used is analytically strong, consistent, and clean as well as that the models are statistically sound and practically useful in real-world medical environments.

The approach comprises several stages. First, the dataset is chosen and filtered to match the aim of forecasting high-cost patients depending on admission-time data. A thorough data cleaning and preprocessing phase then handles missing values, removes anomalies, and standardized varying formats. Techniques of feature engineering and transformation help to improve the predictive ability of the data. Trend, correlation, and variable distribution visual and statistical

evaluations are done using exploratory data analysis (EDA). Class balancing methods are used to avoid biased learning in healthcare cost data since their inherent imbalance causes.

Then developed, trained, and tuned are supervised learning models—regression and classification alike. Explainability and ethical design are especially underlined, which results in the use of interpretable tools such SHAP for feature attribution. To guarantee generalisability and fairness, the last models are rigorously tested with cross-valuation techniques and standard performance criteria.

Beginning with data selection, every level of this approach is covered in the following subsections.

3.2.1 Data Selection

The selected dataset for this study is Hospital Inpatient Discharges data produced by the Statewide Planning and Research Cooperative System (SPARCS) for the state of New York, USA. Established in 1979 thanks to government and industry collaboration, SPARCS is a thorough all-payer data reporting system. Originally, the system was designed to compile data on hospital discharge rates. Currently, SPARCS gathers patient-level data on patient attributes. Comprising patient demographics, diagnosis, treatments, charges, and outcomes, this dataset provides thorough, de-identified records of inpatient hospital discharges throughout New York State. It is a great tool for examining expenses and use of healthcare.

Because of its broad coverage and granularity—variables including age, gender, length of stay, primary diagnosis, and total charges—the SPARCS dataset was selected. These factors are absolutely essential for spotting trends and high healthcare cost predictors. Furthermore, the de-identified character of the data guarantees patient privacy in line with moral research guidelines. The annual updates of the dataset and great sample size help to improve its dependability and applicability for predictive modelling in healthcare economics.

With data across 33 columns, the 21,35,260 row dataset totals roughly 70 million datapoints. By means of this dataset, the thesis seeks to create predictive models capable of spotting patients at risk of incurring significant medical expenses, so guiding resource allocation and policy decisions in healthcare administration.

Following table provides the detail about the variables present in the dataset, their description and the type of data. This table contains variables as defined by CCSR and APR. The description of these two is as follows:

CCSR (Clinical Classifications Software Refined): CCSR groups medical diagnoses into clinically meaningful categories to simplify healthcare data for research and analysis. It focuses solely on the diagnosis without considering severity or complexity of the condition.

APR (All Patient Refined Diagnosis-Related Groups): APR classifies patients based on their diagnosis and the severity of their condition, aiding in risk adjustment and hospital reimbursement. It considers factors like comorbidities and complications to determine appropriate treatment and costs.

Table 3.1 Description of dataset features

Variable	Description	Type
Hospital Service Area	A description of the Health Service Area (HSA) in which the hospital is located. Blank for enhanced de-identification records. Capital/Adirondack, Central NY, Finger Lakes, Hudson Valley, Long Island, New York City, Southern Tier, Western NY.	Text
Hospital County	A description of the county in which the hospital is located.	Text
Operating Certificate Number	The facility Operating Certificate Number as assigned by NYS Department of Health.	Text
Permanent Facility Id	Identifier for the Permanent Facility	Text
Facility Name	The name of the facility where services were performed was based on the Permanent Facility Identifier (PFI), as maintained by the Division of Health Facility Planning.	Text
Age Group	Age of the patient at the discharge categorized into age groups of 0 to 17, 18 to 29, 30 to 49, 50 to 69, and 70 or Older.	Text

Zip Code - 3 digits	The first three digits of the patient's zip code.	Text
Gender	Gender of the patient. F stands for female, M for male and U for unknown	Text
Race	Race of the patient.	Text
Ethnicity	Ethnicity of the patient.	Text
Length of Stay	The total number of patient days at an acute level and/or other than acute care level (excluding leave of absence days).	Text
Type of Admission	A description of the manner in which the patient was admitted to the health care facility: Elective, Emergency, Newborn,Trauma, Urgent.	Text
Patient Disposition	The patient's destination or status upon discharge.	Text
Discharge Year	Year of Discharge	Text
CCSR Diagnosis Code	AHRQ Clinical Classification Software Refined (CCSR) Diagnosis Category Code.	Text
CCSR Diagnosis Description	AHRQ Clinical Classification Software Refined (CCSR) Diagnosis Category description.	Text
CCSR Procedure Code	AHRQ Clinical Classification Software Refined (CCSR) Diagnosis Procedure Code.	Text
CCSR Procedure Description	AHRQ Clinical Classification Software Refined (CCSR) Diagnosis Procedure description.	Text
APR DRG Code	The All Patients Refined Diagnosis Related Groups (APR-DRG) Classification Code.	Text
APR DRG Description	The APR-DRG Classification Code Description in Calendar Year 2021, Version 38 of the APR- DRG Grouper.	Text
APR MDC Code	All Patient Refined Major Diagnostic Category Description.	Text
APR MDC Description	All Patient Refined Major Diagnostic Category (APR MDC) Description.	Text

APR Severity of Illness Code	The APR-DRG Severity of Illness Code	Text
APR Severity of Illness Description	All Patient Refined Severity of Illness (APR SOI) Description.	Text
APR Risk of Mortality	All Patient Refined Risk of Mortality (APR ROM) Description.	Text
APR Medical Surgical Description	The APR-DRG specific classification of Medical, Surgical or Not Applicable.	Text
Payment Typology 1	A description of the type of payment for this occurrence.	Text
Payment Typology 2	A description of the type of payment for this occurrence.	Text
Payment Typology 3	A description of the type of payment for this occurrence.	Text
Birth Weight	The neonate birth weight in grams; rounded to nearest 100 g.	Numeric
Emergency Department Indicator	Parameter to identify whether a patient's visit or admission occurred through the emergency department (ED) of a hospital	Text
Total Costs	Total estimated cost for the discharge.	Numeric
Total Charges	Total charges for the discharge.	Numeric

3.2.2 Data Pre-processing

Particularly in the healthcare sector where data is sometimes contaminated by inconsistencies, missing entries, and non-standardized formats, data preprocessing is among the most important stages in any machine learning pipeline. Healthcare records show a complicated interaction of human behaviour, clinical judgement, and administrative coding practices unlike synthetic datasets or structured financial databases. Therefore, the raw data must be cleaned, validated, and reformatted to assure it is appropriate for analytical purposes before any model can be trained.

This work pre-processes using a methodical, multi-step approach to guarantee the consistency, completeness, and quality of the dataset used for predictive modelling. Without leaking any

information from the target variable into the predictors, the emphasis is on building a trustworthy and ethically usable dataset for cost prediction and classification activities.

Data Cleaning

Eliminating oddities and discrepancies in the raw data comes first. Typographical mistakes, invalid timestamps, duplicate rows, and biologically unrealistically high values abound in hospital records. Such entries compromise results and seriously skew model development.

In this sense, data cleaning consists in:

- Removing duplicate rows, maybe resulting from repeated entries for the same discharge record.
- Making sure admission and discharge dates line up chronologically.
- Verifying numerical fields—such as age or length of stay—to lie within clinically and biologically reasonable ranges.
- Rule-based filters help to either correct or exclude invalid records.

Maintaining data integrity and making sure the models avoid learning from mistakes depend on these actions.

Handling Missing Values

In healthcare databases, missing data is a major problem particularly in administrative sources where information is entered manually or gathered between several departments. Missing values in this work will be handled with context-appropriate imputation methods:

- Mode imputation for categorical variables such as payer type, gender, or colour where the most often occurring category serves as a proxy.
- Median imputation for often right-skewed continuous variables like cost or length of stay. The median provides a strong estimate free from influence from very extreme outliers.

Sometimes missingness in itself could be useful. A missing procedure code might suggest, for instance, a non-surgical case. Under these circumstances, a placeholder category like "Unknown" is used instead of imposing possibly false values.

Usually more than 30% missing, variables with extreme missingness are examined case-by-case for consolidation or exclusion. In subsequent research, this helps to avoid noise and duplication.

Outlier Detection and Treatment

The heavy-tailed character of medical expenses—especially in emergency care, surgery, or episodes of a critical illness—makes healthcare cost statistics intrinsically skewed. While some high-cost values could be justified, others could be erroneous or rare enough to limit generalisation of model.

To handle outliers:

- Extreme values in numerical fields are found by applying IQR-based filters and box plots.
- Particularly in cases when the value is not invalid but rather rare, domain-driven caps or log transformations are regarded as alternative to removal. For cost data, for example, a log transformation compresses high values without losing their relative significance.

Crucially, outlier treatment is done free from direct reference to the target variable to prevent inadvertent label leakage.

Standardization of Variable Formats

Standardising formats for consistency and compatibility is crucial in a dataset including several variable kinds—string labels, numerical values, binary indicators, and date fields. this covers:

- Using consistent date formats will help to enable accurate computation of duration.
- Turning categorical text into lowercase or consistent label forms helps to avoid duplication.
- For downstream processing libraries, cast variables to suitable data types—e.g., integer, float, category—for compatibility.

Though subtle, these changes are crucial in preventing model mistakes or inconsistencies.

Integrity and Reproducibility Checks

At last, every stage of data preparation is carried out in a repeatable pipeline that can be used on test and training sets alike. This guarantees consistency and helps to prevent data leaking. Traceability and openness are promoted by logging of transformation steps, intermediate summaries, and versioning of cleaned datasets—especially in healthcare research, where repeatability is fundamental of scientific credibility.

All things considered, data preparation in this work is not seen as a peripheral chore but rather as a basic need for ethical, accurate, and deployable machine learning. This stage lays a strong basis for all next modelling activities by guaranteeing a clean, consistent, objective dataset.

3.2.3 Data Transformation

The process of turning structured, pre-processed data into a format more fit for machine learning algorithm analysis is data transformation. Pre-processing fixes mistakes and inconsistencies; transformation concentrates on data reshaping to improve model learning, increase representation, and match algorithmic needs. Careful transformation is crucial in healthcare analytics—where data is heterogeneous, high-dimensional, and often skewed—to guarantee accurate and generalisable model performance.

Key transformation techniques used in the research—feature engineering, encoding, normalisation, and log transformation—are described in this part.

Feature Engineering

Feature engineering is the removal from the dataset of irrelevant variables from the standpoint of the research. Regarding healthcare data, these kinds of events can be rather common since contemporary databases keep many data points for every patient.

Creating new variables from raw data that give the model more instructive signals is another element of feature engineering. Derived variables sometimes provide better predictive value than the raw inputs in the framework of healthcare cost prediction. This covers the development of new variables to gauge relative behaviour of variables, classification of variables into meaningful

bands. Such derived variables can improve model interpretability and lower noise from very variable continuous inputs.

Categories Encoding

Many categorical variables abound in healthcare datasets: diagnosis codes, payer type, patient sex, discharge status, and hospital identifiers. Most machine learning systems cannot use these factors directly; they must be converted into numerical form.

Different type and cardinality lead different encoding techniques to be used:

- Low cardinality categorical variables—such as gender, payer type—have one-hot encoding applied to them whereby every category is turned into a binary column.
- High-cardinality variables such as diagnosis or procedure codes are frequency-based grouped, meaning rare values are consolidated into a "Other" or "Low-frequency" group to lower sparsity and overfitting risk.
- Classifying elements where there are excessively many feature categories. Ignoring this results in high dimensionality in the dataset and might distort the generalisation of the model at large**.**

Standardisation and Scaling

Many times, machine learning techniques presume that input data are on a like scale. While some models, particularly linear models, logistic regression, and support vector machines, perform better when inputs are standardised, others—XGBoost and random forests—are scale-invariant.

- Particularly helpful when model output must be probabilistic or limited, min-max scaling turns features into a 0–1 range.
- With unit variance, Z-score standardising centres variables around zero, so benefiting algorithms sensitive to scale.

Depending on the model's needs, scaling is done on continuous factors including age, cost, and length of stay.

Log transformation

Logarithmic transformation is used to numerical data in order to solve skewness and lower the impact of outliers. Usually right-skewed, these variables have log-transformed equivalents with more normal-like distributions.

Log transformation not only stabilises variance but also increases the interpretability of coefficients in linear models and boosts the performance of algorithms presuming symmetric distributions.

3.2.4 Exploratory Data Analysis

A fundamental part of the research process, exploratory data analysis (EDA) links informed model development with raw data comprehension. Especially in healthcare datasets, which often show skewness, sparsity, and multivariate complexity, EDA helps development of understanding of data distributions, relationships, and anomalies—opposite from simply statistical summaries.

This work uses exploratory data analysis (EDA) not only for descriptive insight but also as an interactive diagnostic tool evaluating data quality, guiding feature selection, and spotting structural patterns pertinent to healthcare cost prediction.

Univariate Analysis

The independent study of every variable to grasp its distribution and range forms the first step of visual analytics:

- Continuous factors including total charges, cost per day, and length of stay are assessed using histograms and density graphs. These graphs enable skewness identification and support logarithmic scaling and other transformations.
- For categorical variables—e.g., gender, payer type, hospital region—bar graphs are used to highlight under-represented subgroups and show dominant categories.
- Outliers are found and demographic segment or cost class variation in distribution examined using boxplots.

Bivariate and Multivariate Exploration

Understanding relationships between pairs or groups of variables takes front stage on the second level of visual analytics. This spans:

- Box Class-based plots allow one to investigate how variables like cost, length of stay, or admission type change across several cost levels.
- Correlation heatmaps identify continuous feature multicollinearity. Strongly linear association variables could be candidates for removal or consolidation to cut duplicity.
- Scatter graphs allow one to investigate interaction between two continuous variables, such length of stay against cost per day or age against cost.

This bivariate study aids in the identification of possible feature interactions relevant for non-linear models such as XGBoost.

3.3 Model Selection: Regression and Classification

Development of a predictive system for high-cost patient identification depends critically on the choice of suitable ML models. Model selection in the healthcare environment has to strike three main balances: performance, interpretability, and feasibility. Two general modelling paradigms—regression and classification—each fit for a different framing of the cost prediction problem are investigated in this work.

Whether the objective is to estimate actual cost values (a regression problem) or to classify patients into risk tiers (a classification problem), the dual approach reflects the need of experimenting with several formulations of the research question.

Regression Modeling

In regression modeling, the objective is to predict the total healthcare cost incurred by a patient during their hospitalization or episode of care. Two regression algorithms were selected:

- **Ridge Regression:** A linear model with L2 regularization, used to mitigate multicollinearity among highly correlated predictors and procedure intensity. Ridge

regression maintains simplicity and interpretability while reducing the risk of overfitting in high-dimensional settings.

- **XGBoost Regression**: A gradient-boosted ensemble method capable of capturing complex, non-linear relationships between variables. XGBoost is particularly effective in healthcare applications due to its ability to handle missing data, incorporate feature interactions, and manage skewed target distributions. The model also offers internal feature importance scores, enabling interpretability in cost driver analysis.

Investigated were regression models to see if continuous cost estimation could produce significant, practical predictions. But performance stability and interpretability were challenged by the highly skewed and erratic character of actual medical costs.

Classification Modeling

To overcome the volatility associated with cost prediction, the study also framed the problem as a classification task, where patients are categorized into predefined cost tiers—such as low, medium, and high. This approach better aligns with how healthcare systems prioritize care management and financial triage in practice.

- **XGBoost Classifier**: Chosen for its robust handling of structured tabular data, XGBoost was applied to perform multi-class classification using features such as diagnosis categories, demographics, cost-normalized indicators, and hospital attributes. It also accommodates class imbalance via parameters like scale_pos_weight, which is particularly useful when the high-cost class is underrepresented.

Classification offers greater interpretability and direct utility in operational decision-making. For example, high-cost risk flags can be used to enroll patients into case management programs, whereas continuous cost estimates may be harder to translate into action.

Comparative Model Framing

While both modeling strategies are implemented, the classification-based approach emerges as a better solution due to:

- Better alignment with real-world decision points, where cost categories are more attainable than raw values.
- Improved model stability across varying data distributions.
- Enhanced model interpretability using SHAP analysis, especially for explaining risk categories.

Nevertheless, regression analysis remains a valuable component of the study. It serves to validate feature relevance, estimate upper-bound costs, and provide a complementary view of how patient attributes scale with healthcare spending.

3.4 Model Evaluation

Understanding the dependability, generalisability, and real-world applicability of a machine learning model depends on its performance evaluation—especially in important fields like cost prediction of healthcare. This work evaluates the performance of both classification and regression models using several evaluation strategies. Key standards in healthcare analytics, accuracy and error minimisation are only one of the metrics selected; also addressed are fairness, robustness, and interpretability.

The performance measures, validation techniques, and model explanation tools applied for assessing the developed models in this study are described in this part.

3.4.1 Evaluation of Regression Models

Regression models were used to estimate continuous healthcare costs. Two core metrics were applied to assess model performance:

- **R^2 Score (Coefficient of Determination):** R^2 Score quantifies the proportion of variance in the dependent variable (e.g., total healthcare cost) that is explained by the independent variables. It provides an indication of model goodness-of-fit.The formula for R^2 Score is

$$R^2 \text{ Score} = 1 - \frac{\Sigma(Y_i - \widehat{Y})^2}{\Sigma(Y_i - \overline{Y})^2}$$

Where Y_i denotes actual value, $\widehat{Y}$ denotes model predicted value and $\overline{Y}$ denotes mean value of all values.

An R^2 value close to 1 indicates that the model explains most of the variance; an R^2 near 0 suggests limited explanatory power.

- **Root Mean Squared Error (RMSE):** RMSE measures the average magnitude of prediction errors. It penalizes large errors more heavily, making it sensitive to outliers—an important factor in skewed healthcare cost data.The formula for RMSE is

$$RMSE = \sqrt{\frac{\Sigma(Y_i - \widehat{Y})^2}{N}}$$

Lower RMSE values indicate better predictive accuracy. RMSE is especially relevant when cost predictions are used for financial planning or budgeting applications.

3.4.2 Evaluation of Classification Models

Classification models were used to segment patients into predefined cost categories (e.g., low, medium, high). The following metrics were used:

- **Accuracy:** Accuracy measures the proportion of total correct predictions across all classes.However, accuracy alone may be misleading in imbalanced datasets—where the high-cost class is underrepresented—necessitating additional metrics.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Where TP denotes True Positives, TN denotes True Negative, FP denotes False Positive and FN denotes False Negative

- **Precision:** Precision evaluates the proportion of true positive predictions among all predicted positives. It is particularly important for controlling false alarms (e.g., incorrectly labeling low-cost patients as high-risk).

$$Precision = \frac{TP}{TP+FP}$$

- **Recall (Sensitivity):** Recall assesses how well the model captures actual high-cost patients. It reflects the model's ability to avoid false negatives—i.e., failing to flag true high-risk cases.

$$\text{Recall} = \frac{TP}{TP+FN}$$

- **F1-Score:** F1-score is the harmonic mean of precision and recall. It is a preferred metric when classes are imbalanced and both false positives and false negatives are costly. A higher F1-score reflects a more balanced classification model.

$$\text{F1-score} = 2 \text{ x } \frac{PRECISION\ X\ RECALL}{PRECISION + RECALL}$$

- **AUC-ROC (Area Under the Receiver Operating Characteristic Curve):** AUC-ROC measures the model's ability to distinguish between classes. It plots the true positive rate (sensitivity) against the false positive rate at various thresholds. An AUC closer to 1.0 indicates excellent class separation; values near 0.5 suggest no better performance than random guessing.

3.4.3 Cross-Validation

K-fold cross-valuation is used to guarantee that the model is not overfitted to the training sample and will generalise well to unobserved data:

- The dataset is split into k subsets (folds)
- The model is trained on k−1k−1 folds and validated on the remaining fold
- The process is repeated k times, each time using a different fold as the validation set
- The average across all folds, so lowering the variance of evaluation, is the final metric.

Especially in cases of moderately sized datasets or highly variable cost profiles, this approach improves the resilience of model evaluation.

3.4.4 Interpretability and Feature Importance (SHAP Analysis)

Given the critical nature of healthcare decisions, model interpretability is a central component of evaluation. To this end, the study utilizes SHAP (SHapley Additive exPlanations), a unified framework that quantifies the contribution of each feature to individual predictions.

SHAP values enable:

- **Global explanation**: Understanding which features most influence model outputs across the entire dataset,
- **Local explanation**: Explaining why a specific patient was classified as high-cost or not,
- **Fairness audits**: Ensuring that sensitive attributes (e.g., gender, race, insurance type) do not unduly bias the predictions.

These insights support transparency, trust, and accountability—critical requirements for deploying ML models in clinical or insurance environments.

3.5 Conclusion

This chapter has detailed the comprehensive methodological framework employed to develop predictive models for identifying high-cost patients using structured healthcare data. It began with data selection and rigorous pre-processing to ensure data quality, followed by transformation techniques tailored to the healthcare domain. Exploratory visual analytics supported deeper understanding of variable distributions and relationships, enabling informed feature engineering.

Subsequently, both regression and classification paradigms were explored, with model selection guided by the nature of the prediction task, interpretability requirements, and practical deployment considerations. The final model evaluation strategy incorporated multiple performance metrics—spanning accuracy, precision, recall, F1-score, AUC, and RMSE—along with robustness checks via cross-validation and explainability through SHAP.

Chapter 4: Exploratory Data Analysis

4.1 Introduction

This chapter focuses on the overview of the dataset that has been chosen for the research along with an analysis of its features. Section 4.2 focuses on the pipeline of dataset preparation which includes selection of relevant variables, transformation of variables,identification and treatment of missing values, and univariate of the variables to derive insights about the variables that are chosen for the analysis. Section 4.3 discusses the bivariate exploratory Data Analysis where we aim to describe the impact of independent variables on the dependent variables on a standalone one-on-one basis and derive impactful insights and check whether these insights present any form of anomaly with respect to the real-world data and find explanation for such anomalies if required. These steps collectively lay the foundation for the model development and further performance evaluation for the prediction and classification model in further chapters.

4.2 Dataset Preparation

The whole data preparation process followed on the raw healthcare dataset before model development is described in this part.Appropriate data preparation becomes essential given the complexity and great number of variables present in the dataset to guarantee that the data become relevant for the model development and deployment. Eliminating variables that had no bearing on the study came first in this level. Identification and treatment of missing values and building pertinent categories for variables with many unique values came next to help to streamline the analysis and model building process.

4.2.1 Elimination of Variables

As a part of the Data pre-processing pipeline, elimination of certain variables was a necessary prerequisite to enhance insights generation, model development and recommendations for the necessary stakeholders. Following table contains the list of variables and the reason due to which the variables were deleted.

Table 4.1 Variables deleted and their Reason

Variable	Reason for elimination
Hospital County	Removed to avoid any unnecessary complexity that may occur as a result of large number of counties.
Operating Certificate Number	Identification Number for Facility and thus of no use for the analysis and research
Permanent Facility Id	Identification Number for Facility and thus of no use for the analysis and research
Facility Name	Name of the Facility and thus of no use for the analysis and research
Patient Disposition	The destination of patient post discharge is irrelevant from the point of view of research
Discharge Year	All the values of the discharge year were 2021 and thus made no significant impact on the research
CCSR Diagnosis Description	Description of the diagnosis
CCSR Procedure Description	Description of the Procedure
APR DRG Code	Identification code for the diagnosis
APR DRG Description	Column had too many values which could have complicated the process of obtaining useful insights and may lead to overfitting
APR MDC Code	Identification code for the Major diagnosis Category
APR Severity of Illness Code	Identification code for the variable
Payment Typology 1	Variable has nothing to do with determining cost
Payment Typology 2	Large number of Missing values (about 51%)
Payment Typology 3	Large number of Missing values (about 84%)

4.2.2 Identification and Treatment of Missing values

Once variables are eliminated, focus shifts to finding columns having NA values and treat them accordingly. In such cases, the NA values are either replaced with appropriate values or the rows are deleted. Following is the result of the number of NA values in each column:

Hospital Service Area	5214
Age Group	0
Zip Code - 3 digits	40246
Gender	0
Race	0
Ethnicity	0
Length of Stay	0
Type of Admission	0
Patient Disposition	0
CCSR Diagnosis Description	0
CCSR Procedure Description	583187
APR DRG Description	0
APR MDC Description	0
APR Severity of Illness Description	589
APR Risk of Mortality	589
APR Medical Surgical Description	0
Payment Typology 1	0
Birth Weight	1925333
Emergency Department Indicator	0
Total Charges	0
Total Costs	0
Residence Area	40246

From the results it can be observed that Primarily 7 columns had NA values: Hospital Service Area, Zip Code- 3 digits, CCSR procedure description, Severity of Illness description, Risk of Mortality and Birth weight. Each of the columns were treated as follows:

Hospital Service Area: Since number of NA values is too small compared to size of dataset (0.25%), the NA values were imputed with the mode value of the array which was New York City. This made sense as about 42% of New York state's population reside in New York city.

CCSR Procedure Description: This column had a significantly high proportion of NA values (~25%). In this case, imputing values through mean or any other method was not possible since the variable relates to a procedure that could have been carried out for a particular diagnosis. As a result, the rows with NA values were removed altogether.

APR Severity of Illness Description: Although the proportion of NA values is very small, (~0.025%), the NA values of this column could not be replaced by mean and hence the row with NA values were removed from the dataset.

APR Risk of Mortality: Although the proportion of NA values is very small, (~0.025%), the NA values of this column could not be replaced by mean and hence the row with NA values were removed from the dataset.

Birth Weight: Since the column had a substantially high number of NA values (~89%), the column was altogether eliminated since imputing values for such a large number of cases could have distorted final results.

4.2.3 Transformation into Categorical Variables

In the dataset, 3 columns, Zip code-3 digits, CCSR Diagnosis description, CCSR Procedure description, were transformed into Categorical variables. Following is the description of each of the transformation:

Zip code-3 digits: The number of unique Zip codes is 50, which could add to the complexities to the research and model development. To reduce this complexity, the Zip codes were mapped to respective counties using New York data and mapping these counties to the division by Hospital Service area to create a Residential Area variable which includes 9 values and provides similar geographical scale as that of Hospital Service Area.

CCSR Diagnosis Code: The number of unique values in Diagnosis description was 478, which is a very high number. As a result, the Code associated with each diagnosis was pulled and mapped to a category of Diagnosis as designated by CCSR which reduced the number of categories to 21. This column is called Diag_Cat.

CCSR Procedure Code: The number of unique values in the procedure description was 321, which is a very high number. As a result, the Code associated with each procedure was pulled and mapped to a category of Procedure as designated by CCSR which reduced the number of categories to 31. This column is called Proc_cat.

4.3 Exploratory Data Analysis

Once the dataset was cleaned and transformed using different methods, univariate and Bivariate analyses were done on the dataset to extract insights and check for any anomaly in the data as opposed to the claims in the existing literature in the field of healthcare costs.

4.3.1 Univariate Analysis

In univariate analysis, the distribution between values was observed to understand the distribution of data. The results according to each variable is as follows:

Hospital Service Area

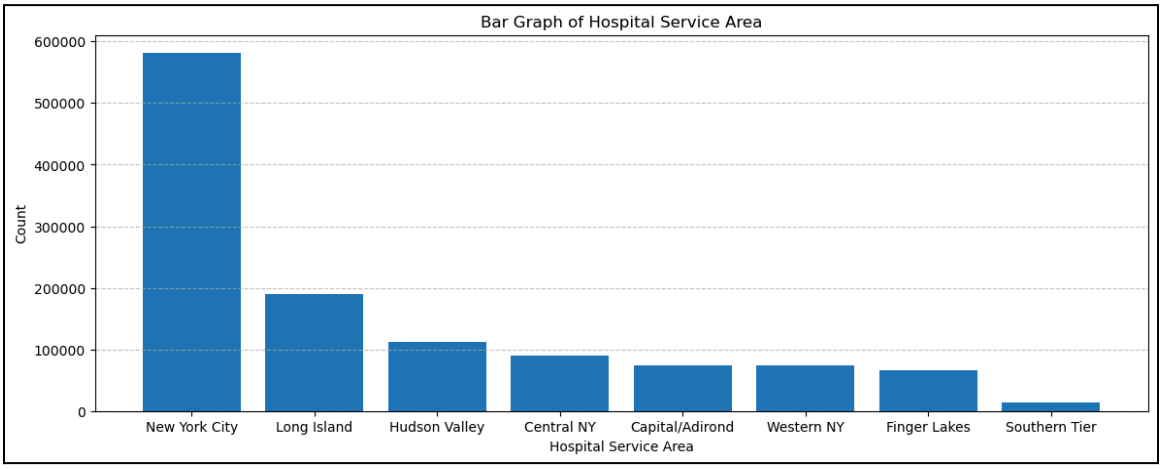

Figure 4.1 Distribution of Hospital Service Area

As shown in figure 4.1, most of the patients discharged were from facilities located in New York City (~48.44%), Long Island (~15.8%), Hudson Valley (~9.33%) and Central New York (~7.46%). These 4 areas combined together account for about 81.03% discharges in the state of New York in that year. These 4 HSAs combined account for about 74% of the entire state

population, pointing towards the fact that the distribution of discharges across HSA is on similar lines of the population distribution

Age Group

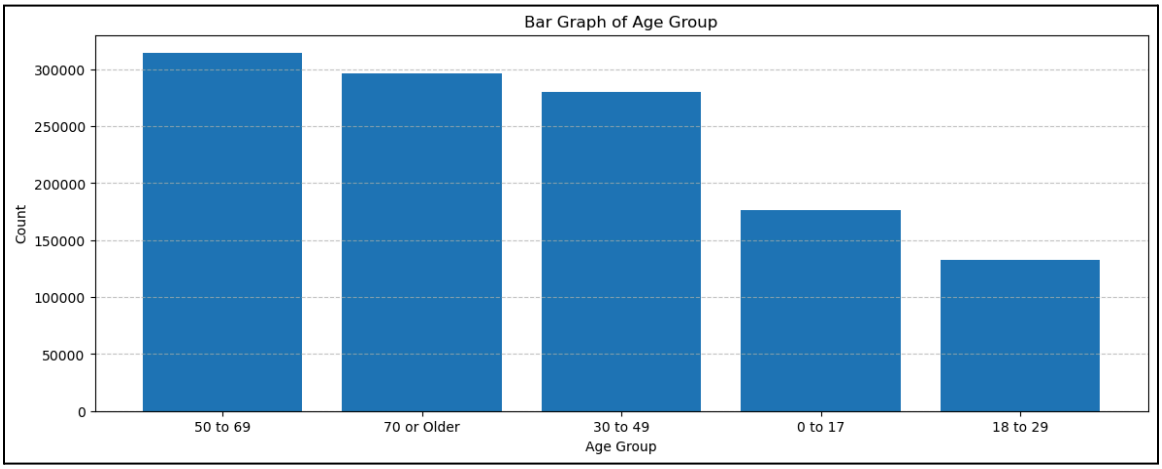

Figure 4.2 Distribution of Age Group

As shown in figure 4.2, the highest number of patients who were discharged were in the age group of 50 to 69 followed by those who were 70 years or older. Combined together, the proportion of patients discharged who are 50 years old or higher is about 51%. Additionally, the number of patients discharged is seen to be increasing as age group increases, only difference being that the age group 18 to 29 years have a lower number that 0 to 17. Combined together, people who are 30 years or older make up about 74.26% of the entire data.

Gender

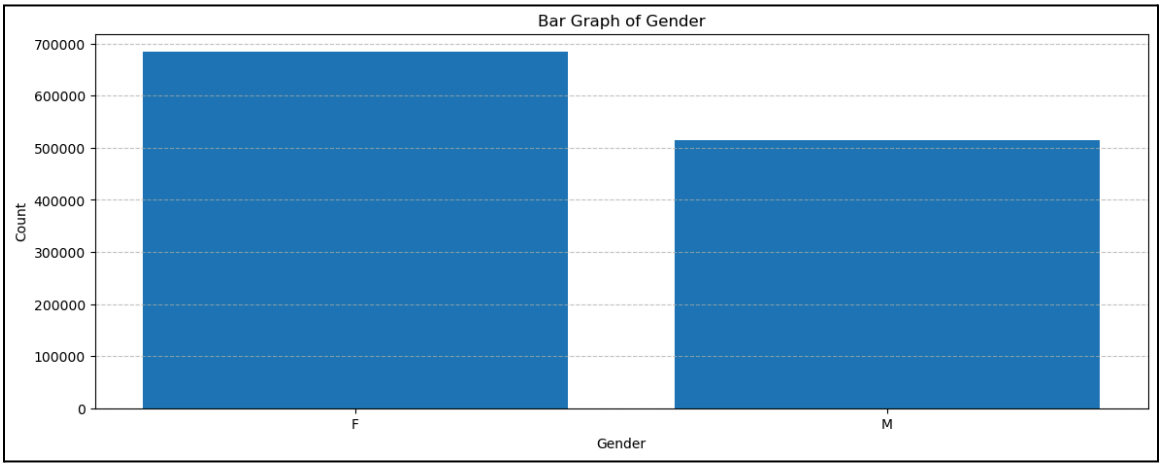

Figure 4.3 Distribution of Gender

As shown in figure 4.3, about 57.04% of the patients are Female and 42.96% of the patients are Male. For the cases where the gender of the patient is unknown, we remove those rows from the dataset since its a very tiny proportion in comparison to the entire dataset.

Race

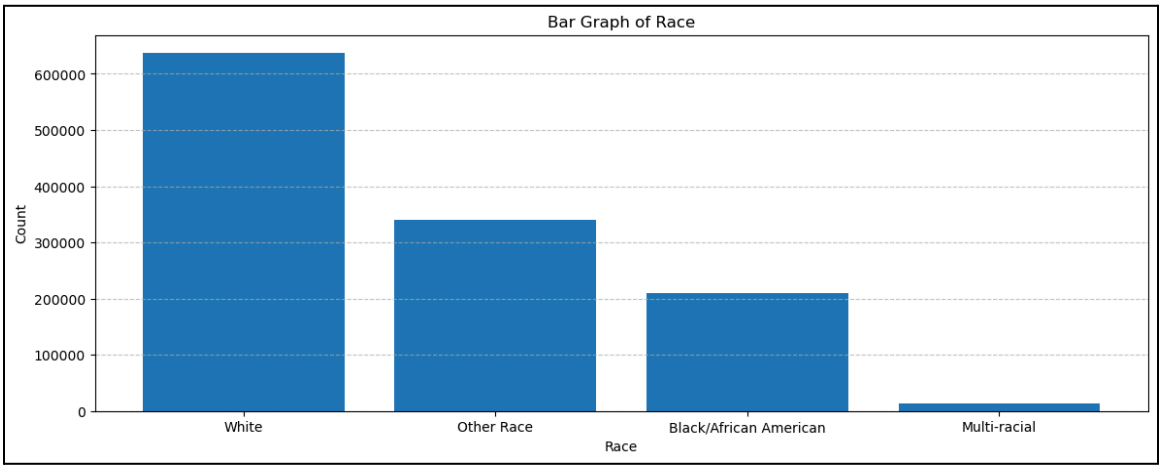

Figure 4.4 Distribution of Race

As shown in figure 4.4, About 53.09% of the patients belong to White alone race while 17.51% patients belong to African-American or Black race. Other races, which could include Asians, Native Americans and other minor races account for 28.33% while Multi-racial people account for a little over 1% of the patients. Here an important point to note is that Hispanic is not considered as a race but an ethnicity by the US Census department.

Ethnicity

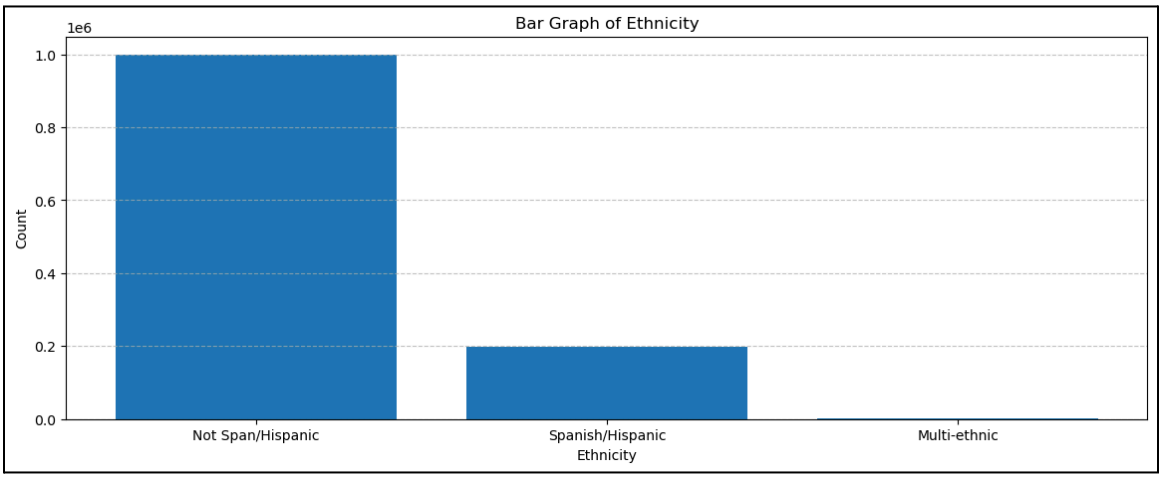

Figure 4.5 Distribution of Ethnicity

As shown in figure 4.5, About 16.56% of the patients are from Spanish/Hispanic Ethnicity while 83.24% are not Spanish/Hispanic. Multi-ethnic people, ie people who belong to more than one ethnicity constitute a tiny 0.2% of the total number of patients

Length of Stay

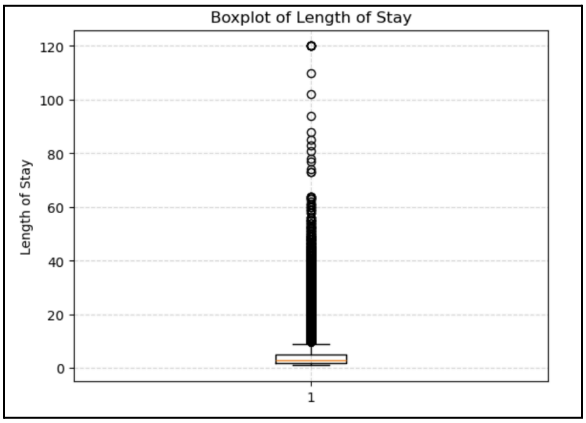

Figure 4.6 Boxplot of Length of Stay

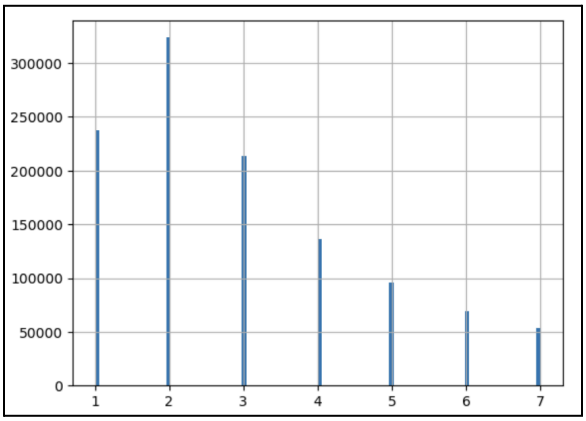

Figure 4.7 Distribution of Length of Stay

As shown in figure 4.6, it can be observed that the median length of stay is 3 days and the first quartile is 2 days and 3rd quartile is 4 days. As a large number of points lie outside of boxplot, it can be observed that the data is highly skewed towards the right and a large number of outliers are present which can possibly influence the entire model and its results. To overcome this problem, all the data points that were either below Q1-1.5 X IQR or above Q3 + 1.5 X IQR were removed. Figure 4.7 shows the distribution of the number of days after removal of outliers and it shows that most patients stayed for 2 days and as length of stay increases, the number of patients decreases.

APR Major Diagnostic category (APR MDC)

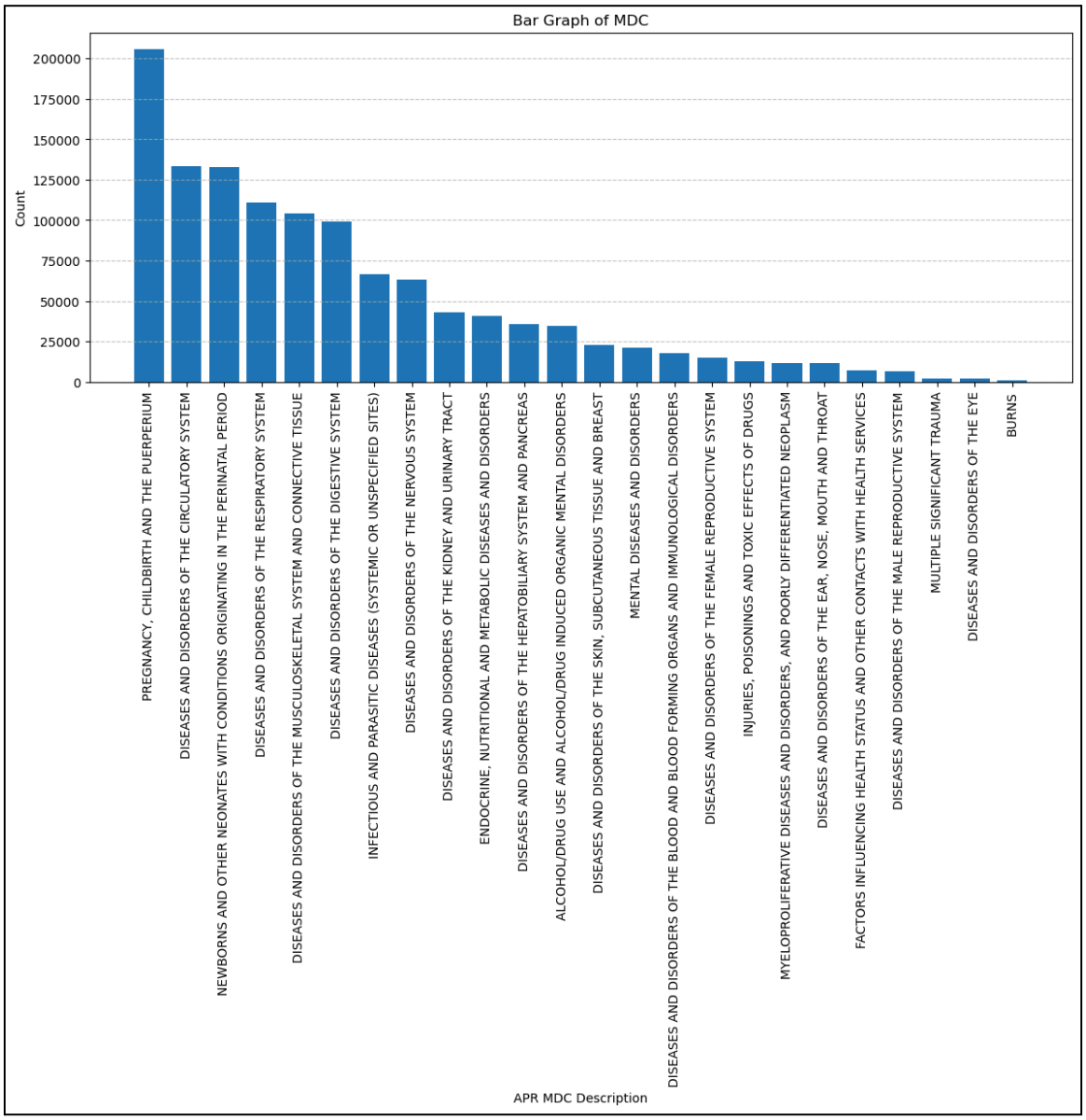

Figure 4.8 Distribution of MDC

As shown in figure 4.9, About 16.6% were admitted for pregnancy, childbirth, and the puerperium, making it the most common reason for hospitalization. This is followed closely by circulatory system disorders and perinatal conditions, accounting for 10.8% and 10.7% of the total patients, respectively. Collectively, these top three categories alone represent around 38.1%of all hospital cases. Other categories include respiratory system disorders (8.9%), musculoskeletal disorders (8.4%), and digestive system issues (8%), suggesting that the majority of hospitalizations stem from physiological or system-level health concerns. Categories such as

infectious diseases (5.4%), nervous system disorders (5.1%), and kidney and urinary tract disorders (3.4%) also contribute notably to the caseload. Rare diagnoses like burns, eye disorders, and multiple significant trauma each account for less than 1% of total cases. Patients with mental health and substance use-related disorders are about 4.5%.

Type of Admission

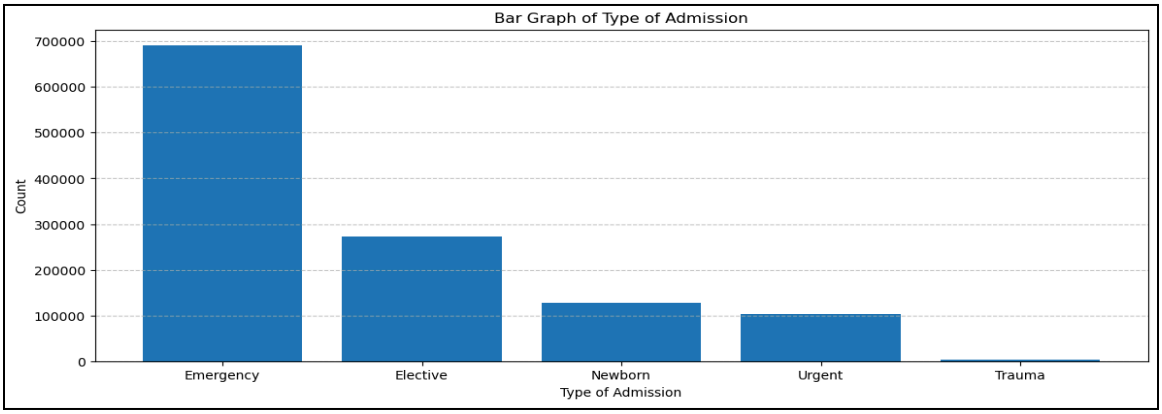

Figure 4.9 Distribution of Type of Admission

From figure 4.9, it can be observed that about 57.57% of patients were admitted in emergency while 20.22% of the patients chose to get admitted for the treatment. 10.70% of patients were admitted for child birth while 8.66% required urgent admission and 2.85% were trauma cases.

APR Severity of Illness

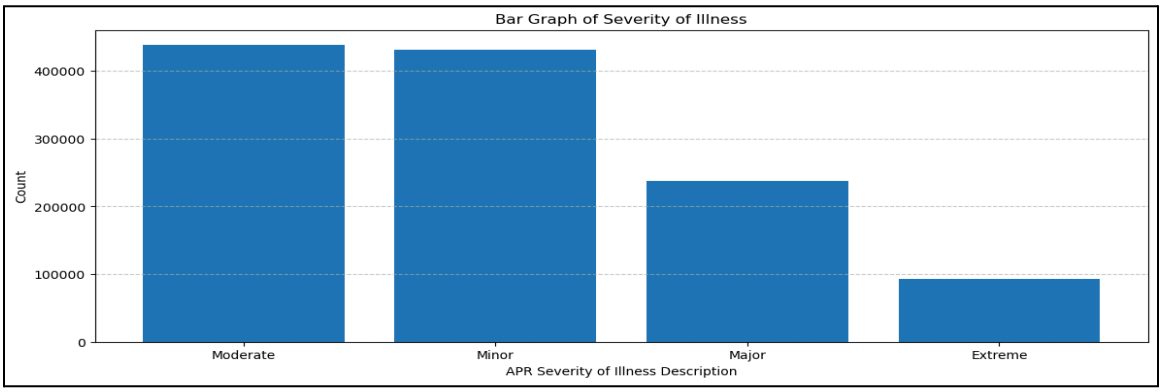

Figure 4.10 Distribution of Severity of Illness

From figure 4.10, it can be observed that about 36.54% of the patients were suffering from Moderate illness while 35.97% were suffering from Minor illness. About 7.71% were suffering

from extreme illness while 19.78% suffered from major illness. From the distribution, it can be observed that about 64% of the patients suffer from moderate to extreme illness.

APR Risk of Mortality

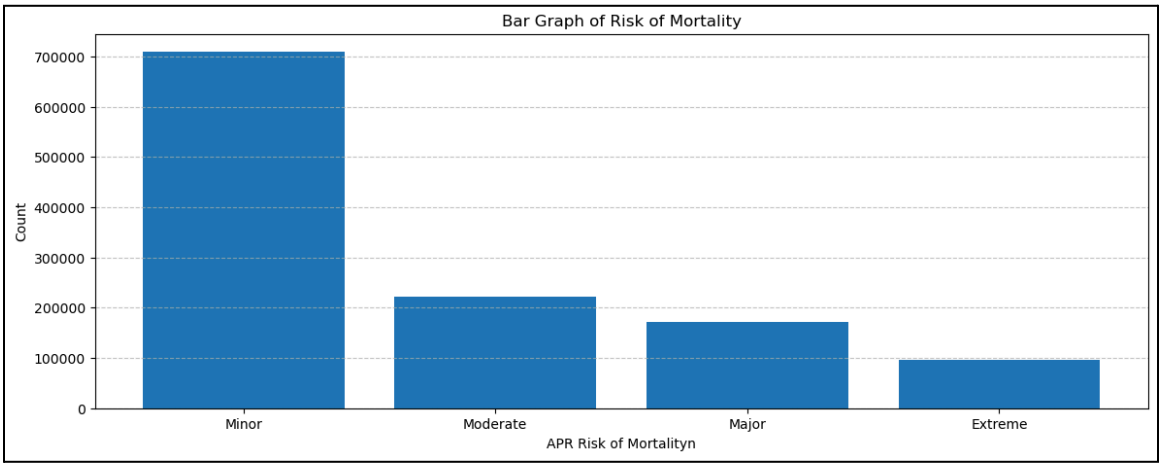

Figure 4.11 Distribution of Risk of Mortality

From figure 4.11, it can be observed that about 58.2% of the patients were at minor risk of mortality followed by patients at Moderate risk (18.2%) and major risk (14.10%). The extreme risk category, accounted for 7.90% of all the patients. From the distribution it can be observed that more than 40% of the patients were at Moderate to extreme risk of mortality.

APR Medical Surgical Description

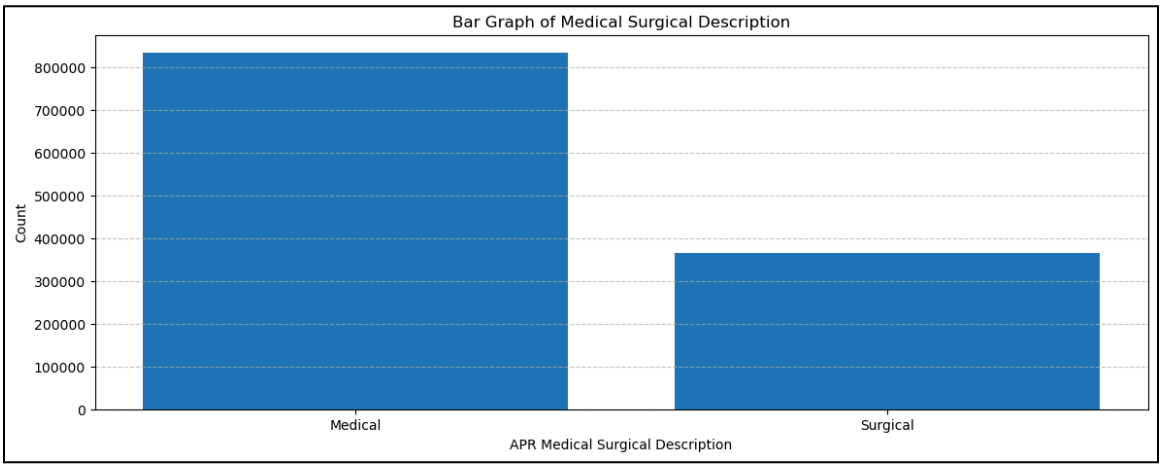

Figure 4.12 Distribution of Medical Surgical Description

From figure 4.12, it can be observed that about 69.5% of the patients were treated for medical conditions that did not require surgery while 30.50% of the patients required surgery. The

dominance of non-surgical medical admissions over surgical admissions showcase the prevalence of conditions that require medical management over operative intervention.

Emergency Department Indicator

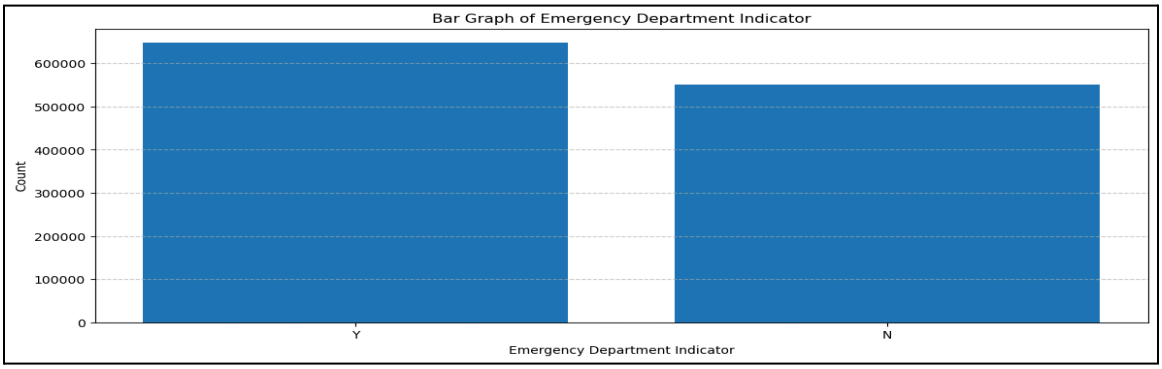

Figure 4.13 Distribution of Emergency Department Indicator

From figure 4.13, it can be observed that about 54% of the patients were admitted through the emergency department while 46% were admitted through non-emergency mediums. This close-split of 8% showcases the critical role of emergency services in the healthcare system.

Residence Area

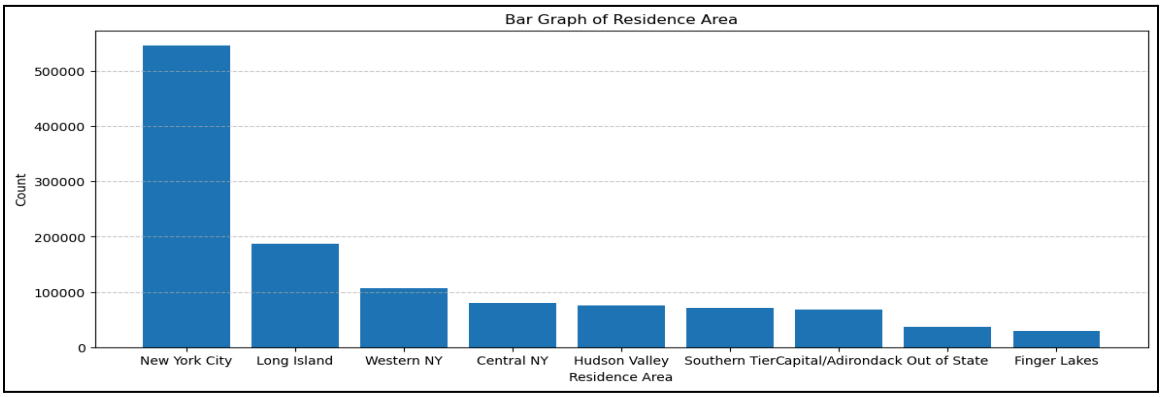

Figure 4.14 Distribution of Residence Area

From figure 4.14, it can be observed that about 45.5% of the patients resided in New York City. This is followed by Long Island (15.6%), Western New York (8.9%), Central NY (6.7%), and Hudson Valley (6.4%). Southern Tier (5.9%) and Capital/Adirondack (5.6%) regions each contribute moderately. Out of state patients are about 3% while Finger Lakes residents account for only 2.4%. The share of Patients in New York City and Long Island (61.1%) shows the heavy demand in urban areas.

CCSR Diagnosis Category

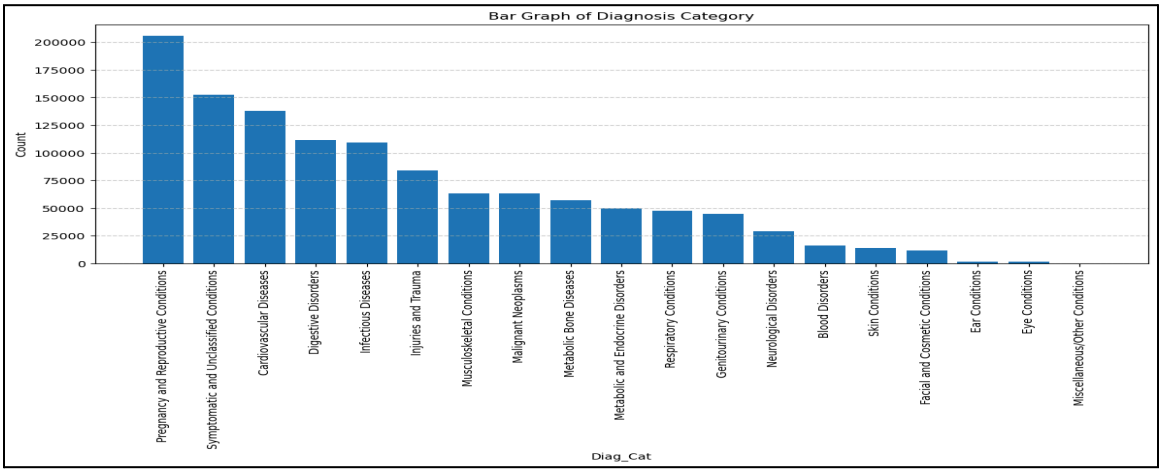

Figure 4.15 Distribution of Diagnosis Category

As shown in figure 4.15, Pregnancy and Reproductive Conditions account for approximately 17% of all cases, followed by Symptomatic and Unclassified Conditions (12.6%), Cardiovascular Diseases (11.4%), and Digestive Disorders (9.2%). Infectious Diseases contribute around 9%, while Injuries and Trauma represent 7%. Moderate proportions are seen in categories like Musculoskeletal Conditions (5.2%), Malignant Neoplasms (5.2%), and Metabolic Bone Diseases (4.7%). Less common categories include Neurological Disorders (2.4%), Blood Disorders (1.3%), and Skin Conditions (1.2%), with Eye and Ear Conditions contributing less than 0.2% each. Miscellaneous or Other Conditions make up a negligible share. Overall, the data shows that top 5 categories account for 59% of the total cases.

CCSR Procedure Category

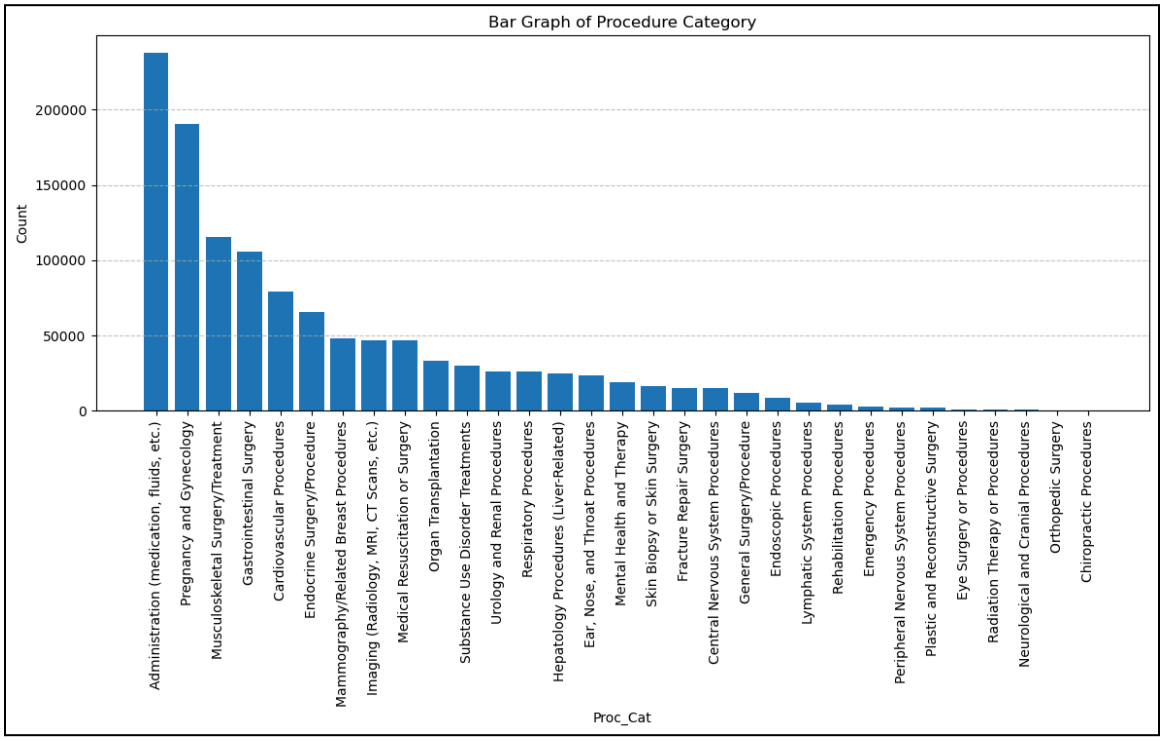

Figure 4.16 Distribution of Procedure Category

As shown in figure 4.16, administration-related procedures make up nearly 20%, followed by pregnancy and gynecology (15.9%), and musculoskeletal surgeries (9.6%). Gastrointestinal and cardiovascular procedures also account for substantial portions, around 8.8% and 6.6% respectively. Categories like endocrine surgeries, breast-related procedures, and imaging each contribute between 4–5%, reflecting common diagnostic and therapeutic activities. Meanwhile, less frequent interventions such as mental health therapy, skin procedures, and fracture repairs form a modest share, and rare categories like neurological, orthopedic, and chiropractic procedures collectively represent less than 1%. This suggests that a relatively small number of procedure types dominate overall hospital care, with the majority of categories occurring far less often.

Total Costs

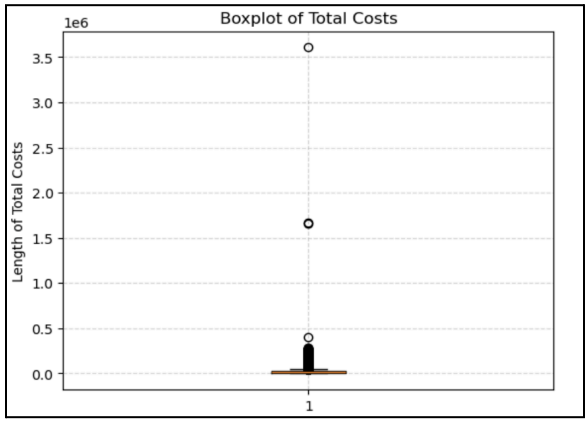

Figure 4.17 Boxplot of Total Estimated Cost

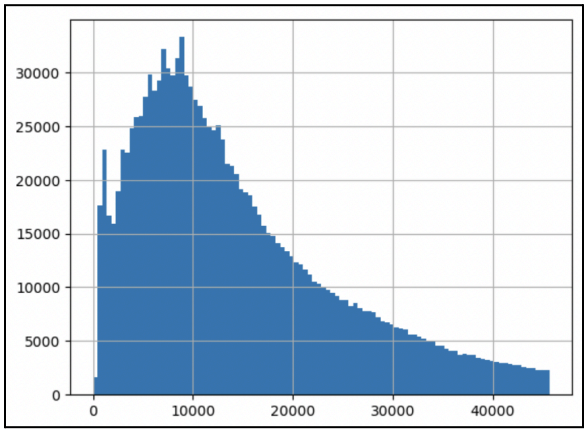

Figure 4.18 Plot of Total Estimated Cost

As shown in figure 4.17, it can be observed that the first quartile is around $7,088 while third quartile is around $22,520.52 quartile while the median cost is around $10,459. As a large number of points lie outside of boxplot, it can be observed that the data is highly skewed towards the right and a large number of outliers are present. To overcome this problem, all the outliers were removed to smoothen the data. Figure 4.18 shows the distribution of the total estimated cost after removal of outliers and it shows that the number of patients increases as estimated cost rises following which the number falls as estimated cost is further increased.

Total Charges

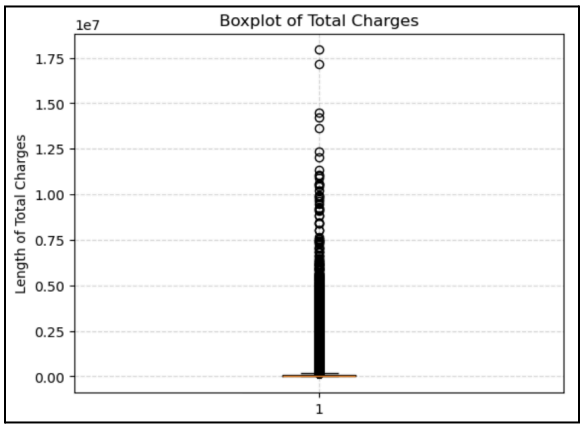

Figure 4.19 Boxplot of Total Charges

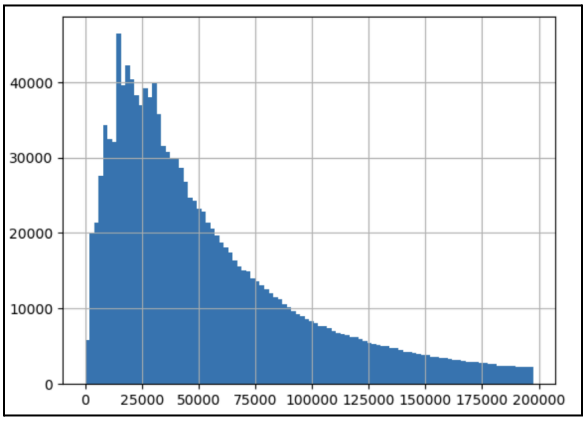

Figure 4.20 Plot of Total Charges

As shown in figure 4.19, the first quartile is around $23,673.37 while the third quartile is around $93,144.625 quartile while the median cost is around $34,172.78. As a large number of points lie

outside of boxplot, it can be observed that the data is highly skewed towards the right and a large number of outliers are present. To overcome this problem, all the outliers were removed to smoothen the data. Figure 4.20 shows the distribution of the Total charges after removal of outliers and it shows that the number of patients increases as Charges rises till a level following which the number falls as estimated cost is further increased.

4.3.2 Bivariate Analysis

In Bivariate Analysis, we try to find the relationship between the independent variables and the target variable (Total Charges). Since the number of observations of the dataset is high, we use Average Charges as a parameter and map categories of each variable against it to draw insights. Results are as follows:

Hospital Service Area

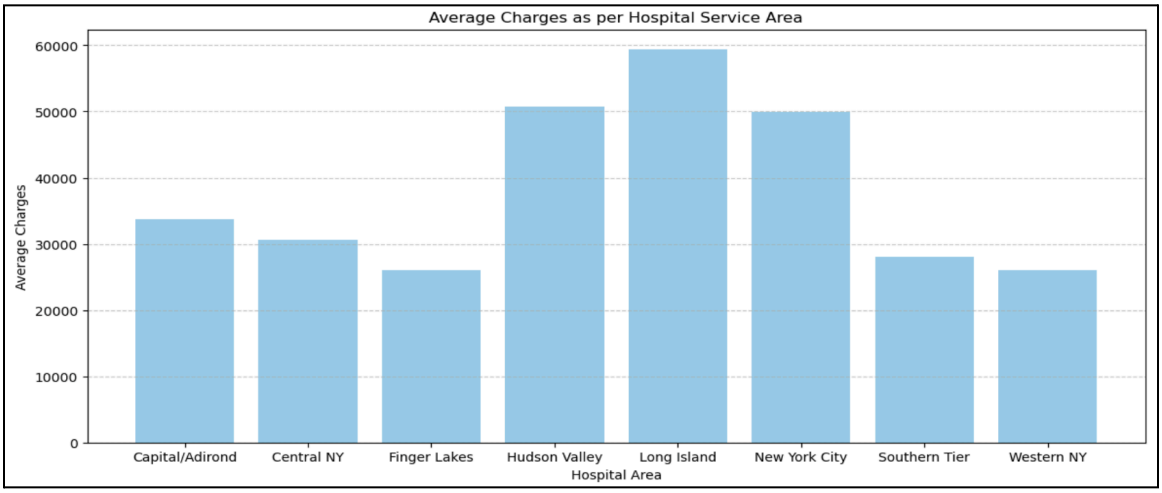

Figure 4.21 Average Charges as per Hospital Service Area

As shown in figure 4.21, Hospitals in the Long Island area have the highest average charges which is followed by Hudson Valley and New York City at the same level of average charges. Grouped together we see that hospitals in these 3 areas charge more than $50,000 on average, showcasing the high healthcare costs in urban and suburban areas. On the other hand, Finger Lakes, Western NY, and Southern Tier have the lowest average charges, generally falling below $30,000. Regions like Capital/Adirond and Central New York lie in the mid-range.

Age Group

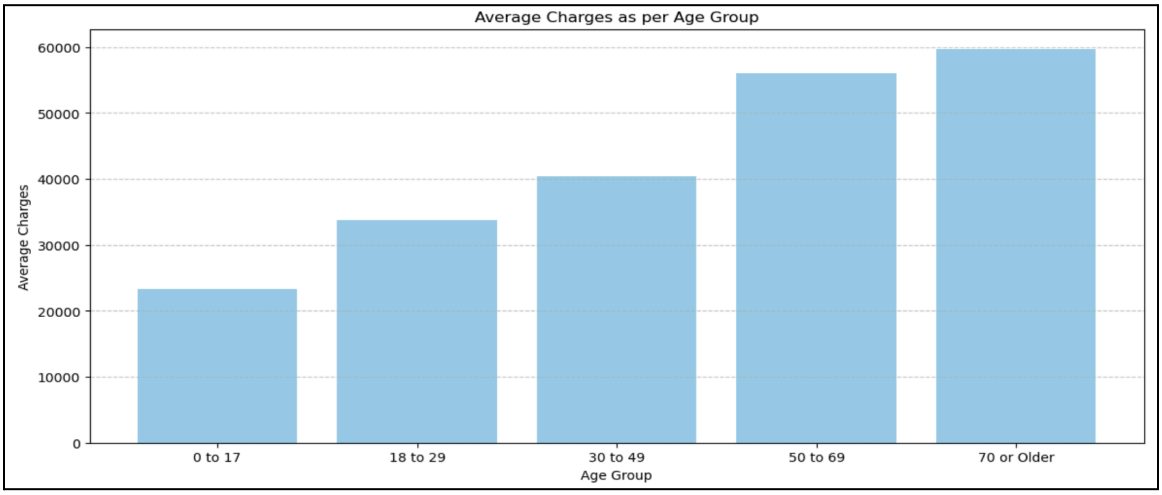

Figure 4.22 Average Charges as per Age group

As shown in figure 4.22, as the age group progresses, the average group increases, signalling a strong positive relationship between the age group of a person and the charges incurred. From this, it can be inferred that if a patient is from an older age group, they may incur higher charges as compared to patients from a younger age group.

Gender

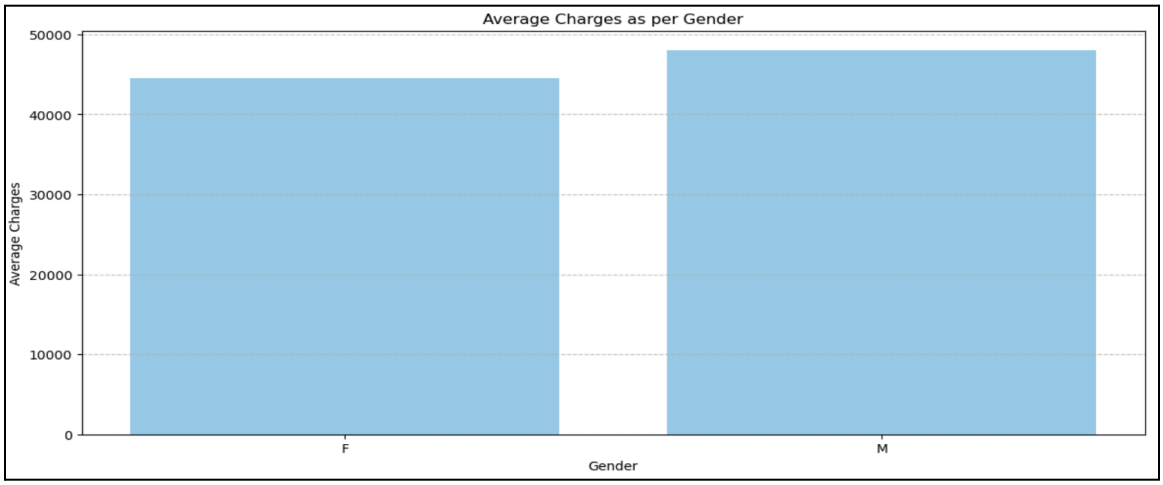

Figure 4.23 Average Charges as per Gender

As shown in figure 4.23, male patients are incurring more charges than female patients and the differential between average charges is about 7%, i.e. males patients are paying about 7% more than female patients.

Race

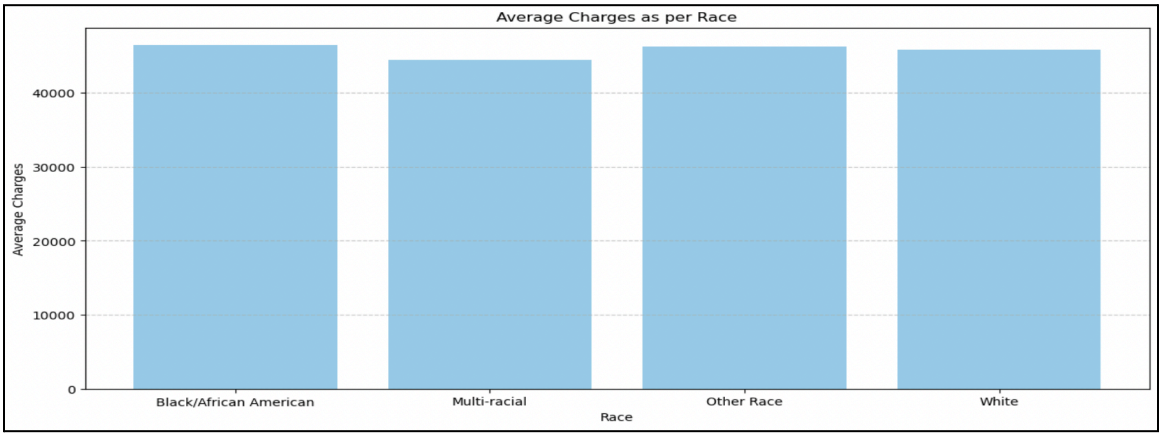

Figure 4.24 Average Charges as per Race

As shown in figure 4.24, there is a minimal difference between average charges that were incurred by different race groups. Black/African American people incur the highest average cost of about $ 46,484 followed by people from other races ($46,195), white ($45,847) and multi-racial people ($44,462). Overall, the graph suggests that the average healthcare costs are fairly consistent across different races.

Ethnicity

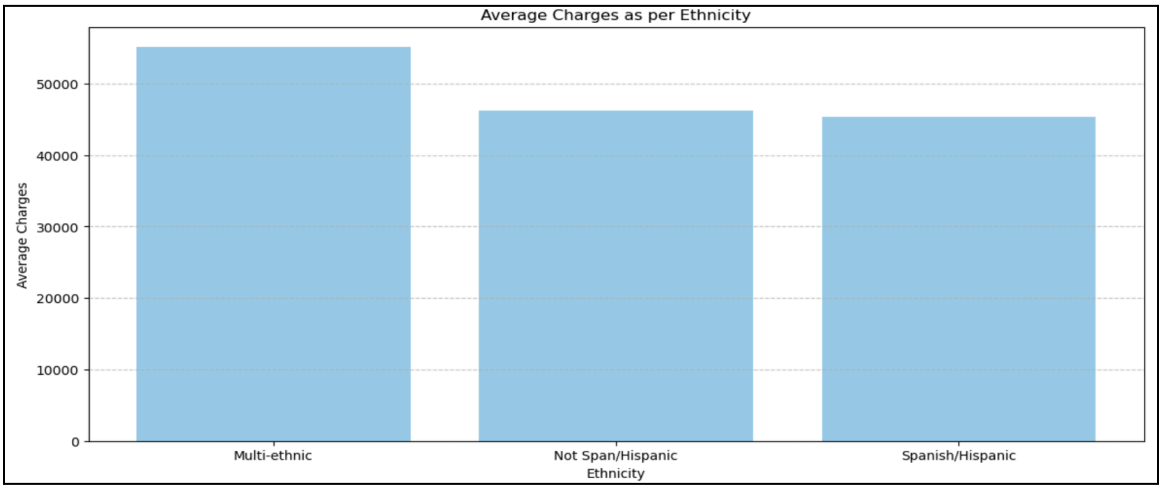

Figure 4.25 Average Charges as per Ethnicity

As shown in figure 4.25, Multi-ethnic people have the highest average group amongst different ethnic groups at $55,159. This is followed by Not Spanish/Hispanic patients ($46,161) and Spanish/Hispanic people having the least average cost at $45,339.

Length of Stay

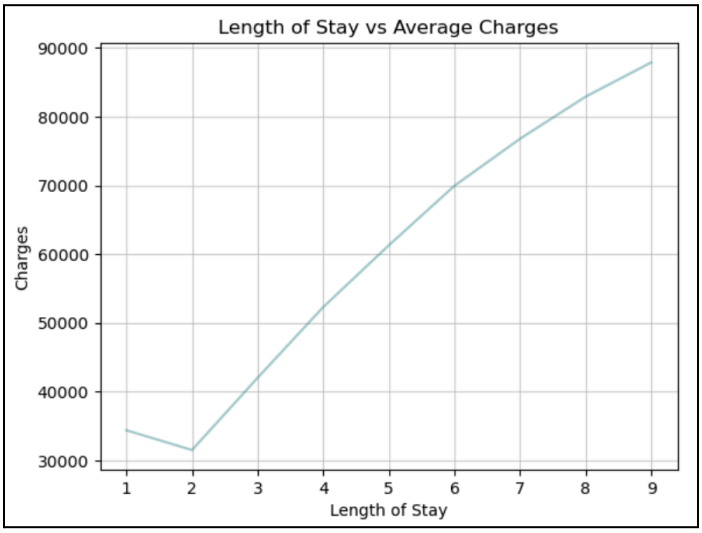

Figure 4.26 Average Charges as per Length of Stay

As shown in figure 4.26, there is a positive linear relationship between length of stay and Average hospital charges. Although there is a dip in average charges at 2, it starts to rise from 3 and continues steadily till 9.The coefficient of correlation between 2 is 0.45, further supporting the positive relationship between the 2 variables.

Type of Admission

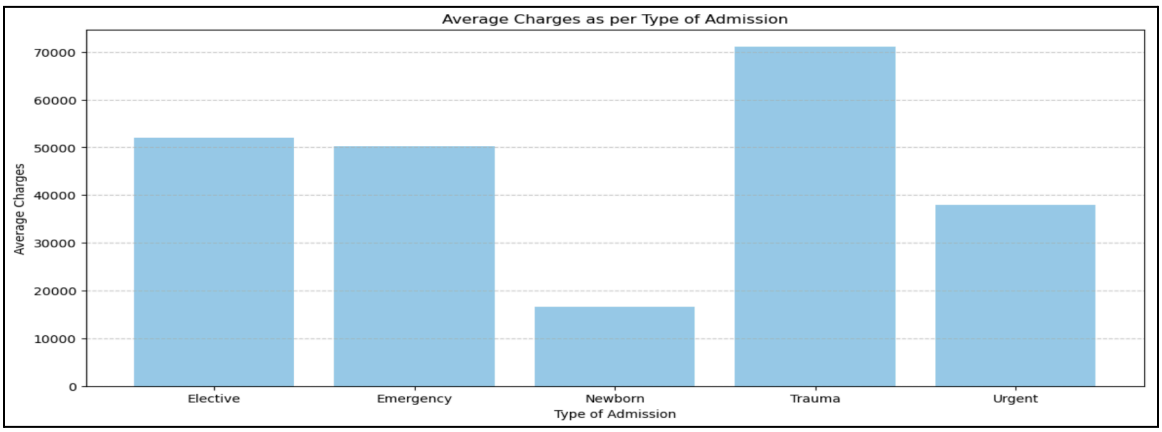

Figure 4.27 Average Charges as per Type of Admission

As shown in figure 4.27, there is a significant difference between average costs incurred for different types of admissions. For Trauma cases, average cost is the highest at $71,145 followed by Elective admissions ($52,082) and Emergency admits ($50,186). For urgent admissions the average cost is $37,955 while admissions related to newborn deliveries have the lowest average

cost at $16,666. This shows that the healthcare cost can vary significantly depending upon what type of case the patient is admitted to the hospital.

APR Major Diagnostic category (APR MDC)

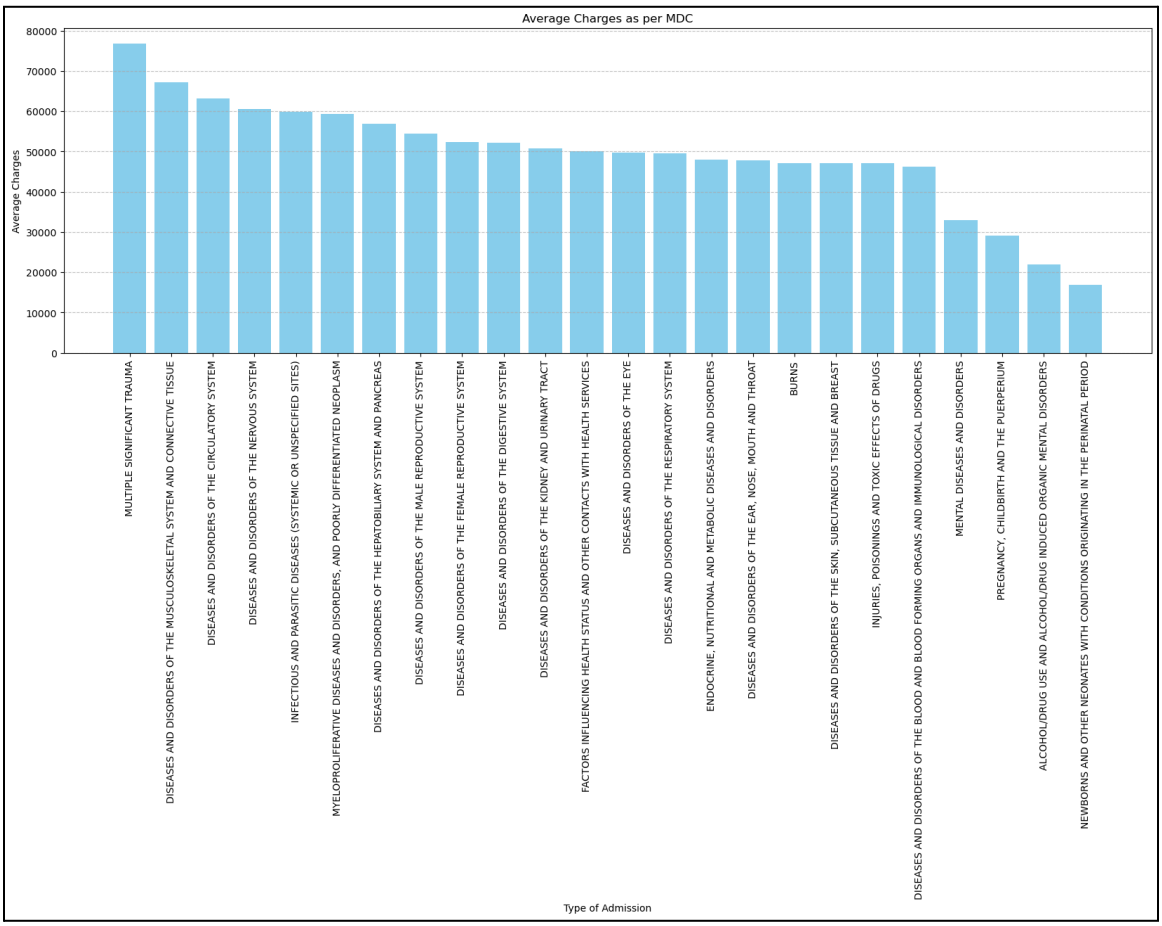

Figure 4.28 Average Charges as per MDC

As shown in figure 4.28, Multiple Significant Trauma leads with the highest average cost at $76,871, followed by Musculoskeletal Disorders ($67,171) and Circulatory System Disorders ($63,214)—all indicating high-complexity, resource-intensive cases. Categories such as Nervous System, Infectious Diseases, and Neoplasms also have elevated costs, ranging between $59,000–$61,000. In contrast, Mental Health Disorders, Pregnancy and Childbirth, Substance Abuse, and Neonatal Conditions reflect the lowest average charges, falling below $33,000, with neonatal care being the least costly at just $16,809. This variation reflects the differing intensity, duration, and clinical complexity across diagnostic categories.

APR Severity of Illness

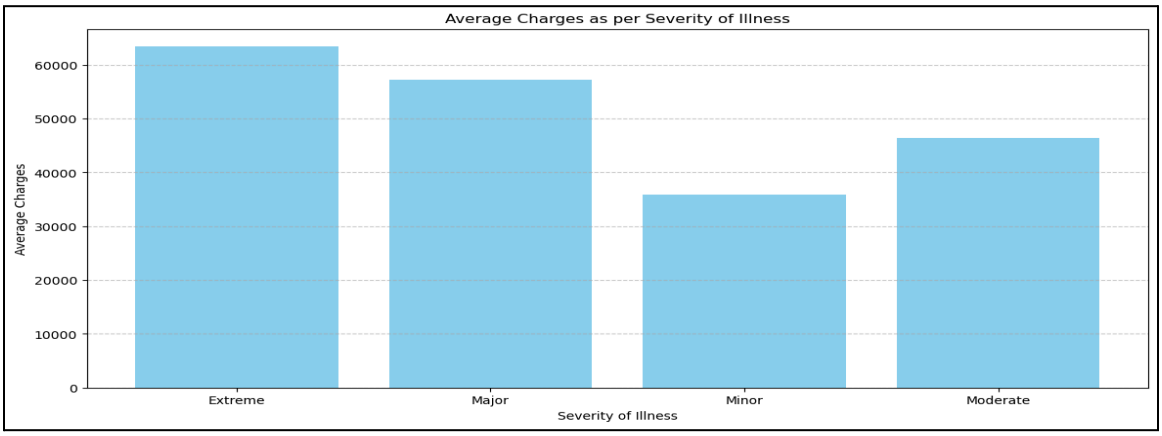

Figure 4.29 Average Charges as perSeverity of Illness

As shown in figure 4.29, cases with extreme illness, the average cost is the highest at $63.466 followed by Major illness ($57,241) and Moderate illness ($46,361). Minor illnesses have the least average healthcare costs at $35,826. On the larger level, the graph shows that the severity of illness can have a significant impact on healthcare costs.

APR Risk of Mortality

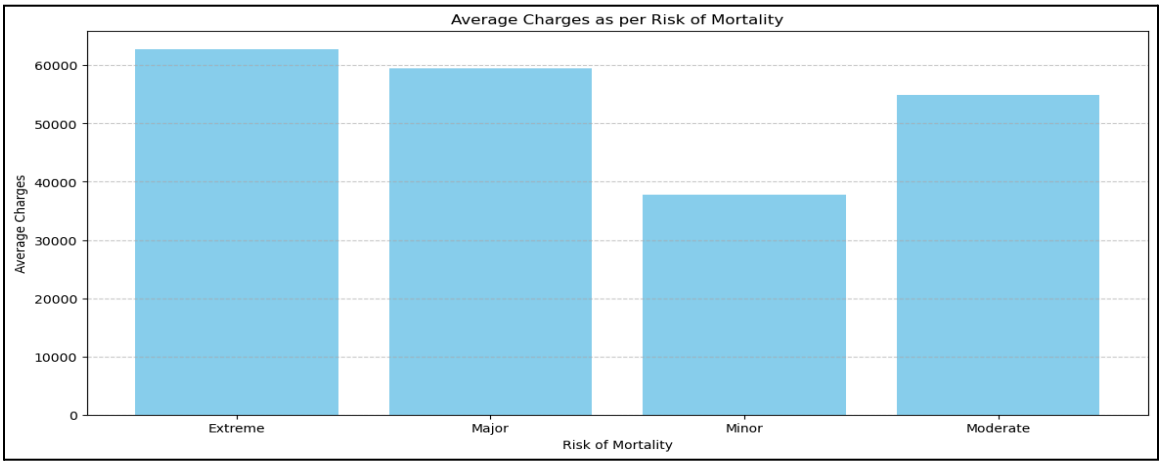

Figure 4.30 Average Charges as per Risk of Mortality

As shown in figure 4.30, cases with extreme risk of mortality, the average cost is the highest at $62.804 followed by Major risk ($59,428) and Moderate risk of mortality ($54,945). Minor risk have the least average healthcare costs at $37,724. On the larger level, the graph shows that the risk of mortality can have a significant impact on healthcare costs.

APR Medical Surgical Description

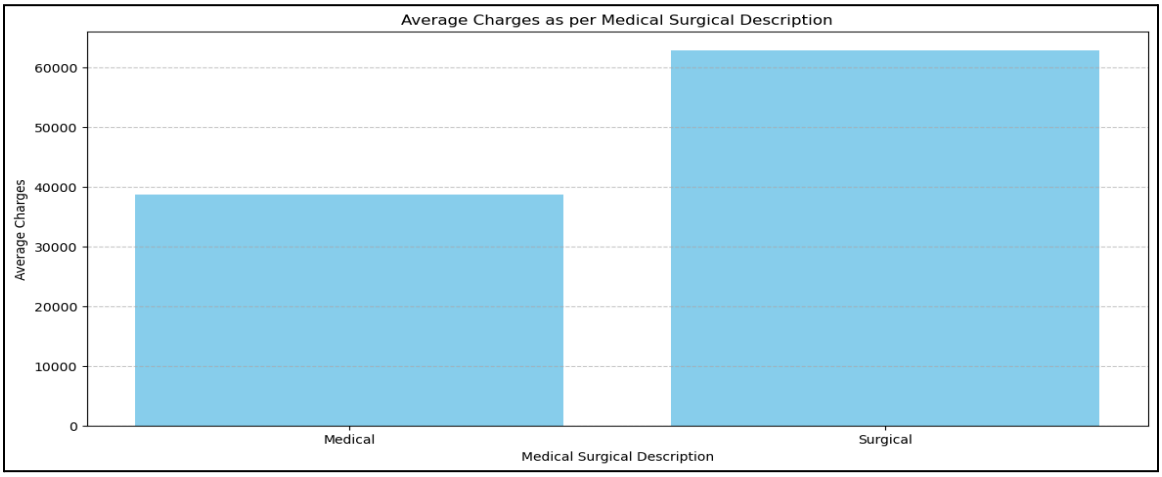

Figure 4.31 Average Charges as per Medical Surgical Description

From figure 4.31, it can be observed that surgical medical admissions had a higher average cost at $62,845 while non-surgical medical admissions have average cost at $38,668.

Emergency Department Indicator

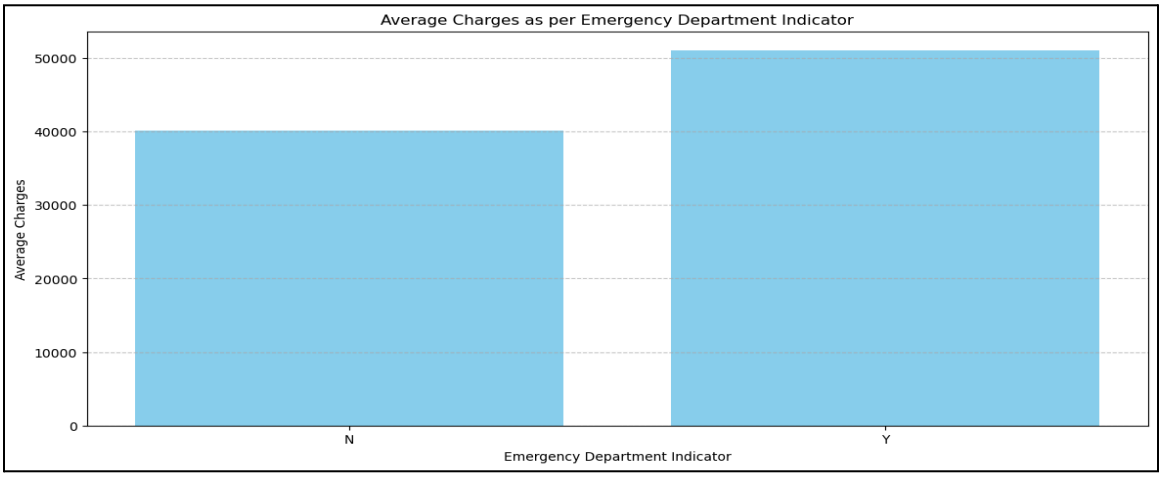

Figure 4.32 Average Charges as per Emergency Department Indicator

From figure 4.32 it can be inferred that admissions through the emergency department had higher average cost at $51,080 while those that were not through the emergency department were at $40,127.

Resident Area

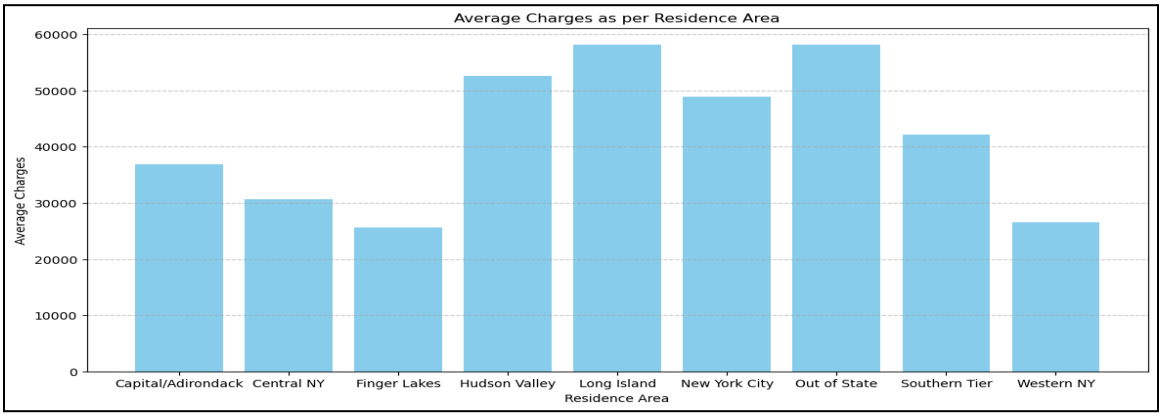

Figure 4.33 Average Charges as per Resident Area

Figure 4.33 shows geographical disparities in the hospital costs for patients. Long Island ($58,211) and Out-of-state ($58,201) are at the top with average costs exceeding $58,000. This is followed by Hudson Valley ($52,614) and New York City ($48,929). These figures suggest that patients from or admitted in more urbanized or affluent regions tend to incur higher healthcare costs. On the other end, residents of Finger Lakes ($25,608), Western NY ($26,621), and Central NY ($30,680) experience significantly lower average charges.

CCSR Diagnosis Category

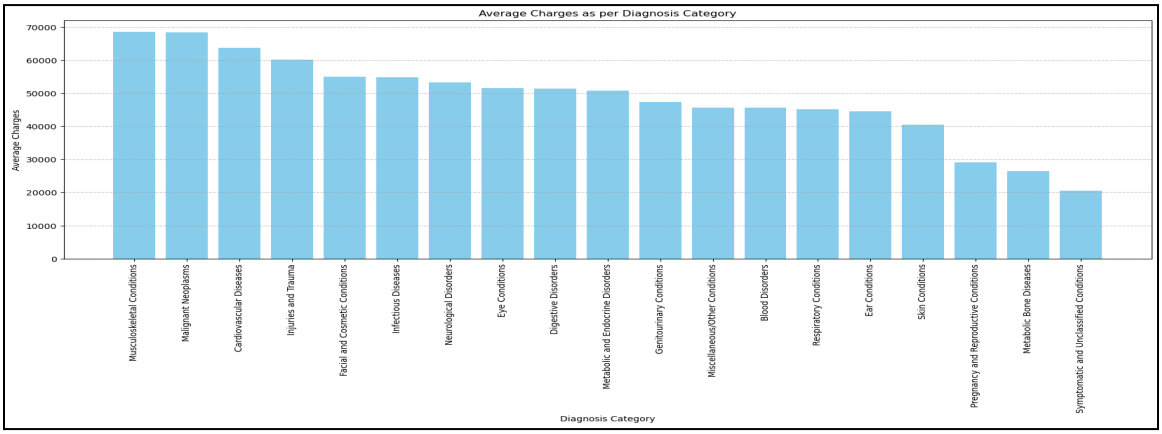

Figure 4.34 Average Charges as per Diagnosis Category

Figure 4.34 shows a stark disparity in average costs across different diagnosis categories. Musculoskeletal Conditions and Malignant Neoplasms top the list, each having average charges of above $68,000, followed by Cardiovascular Diseases ($63,768) and Injuries and Trauma ($60,138). Mid-range categories like Facial and Cosmetic Conditions, Infectious Diseases,

Neurological Disorders, and Digestive Disorders fall between $51,000 to $55,000. On the lower end, Skin Conditions, Pregnancy and Reproductive Conditions, Metabolic Bone Diseases, and Symptomatic and Unclassified Conditions show the least financial burden, with the latter being the lowest at just over $20,000.

CCSR Procedure Category

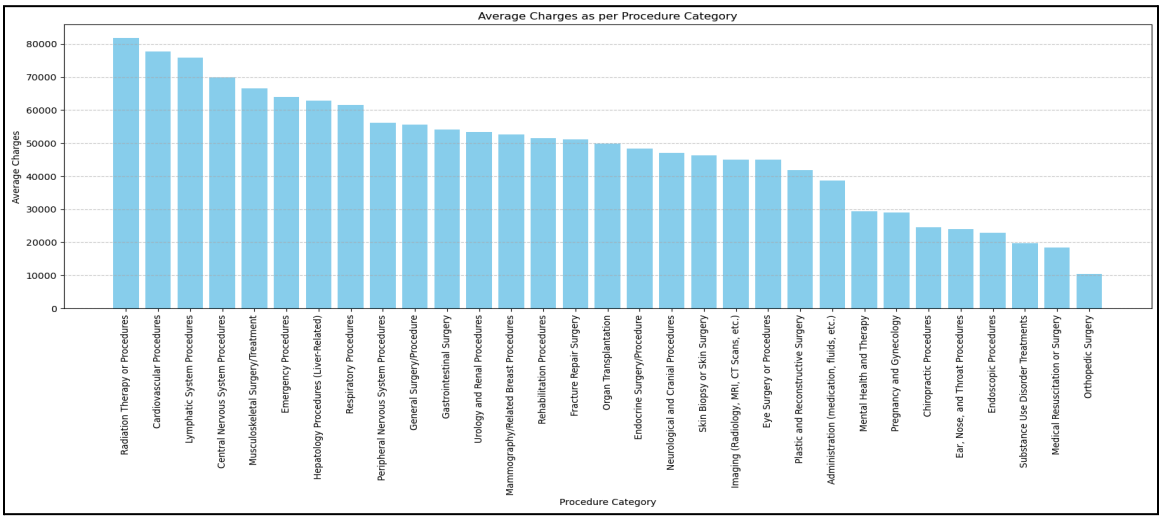

Figure 4.35 Average Charges as per Procedure Category

Figure 4.35 shows a stark disparity of average cost across different categories of procedures. Radiation Therapy ($81,786), Cardiovascular Procedures ($77,732), and Lymphatic System Procedures ($75,785) top the list, reflecting the high cost and complexity of these treatments. Also notable are Central Nervous System and Musculoskeletal Procedures, each averaging over $66,000. Mid-tier procedures like General Surgery, Gastrointestinal Surgery, and Urology Procedures fall between $53,000–$56,000, while Organ Transplantation and Endocrine Surgery range around $48,000–$50,000. On the lower end, categories such as Mental Health and Therapy, Pregnancy and Gynecology, and Substance Use Disorder Treatments have much lower costs, with Orthopedic Surgery being the least expensive at $10,346.

Total Costs

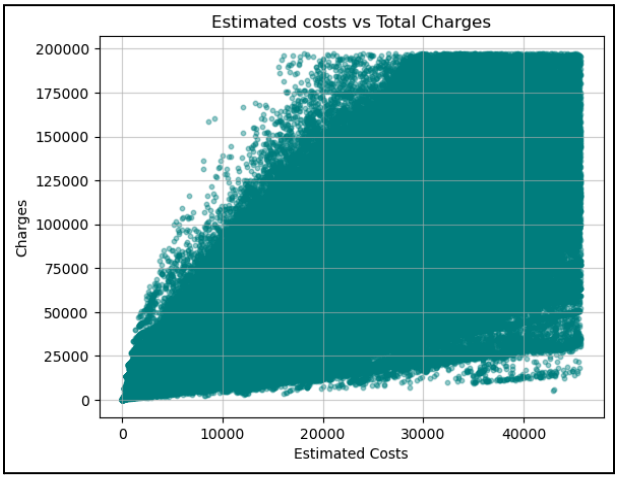

Figure 4.36 Scatter Plot between Estimated cost and Total Charges

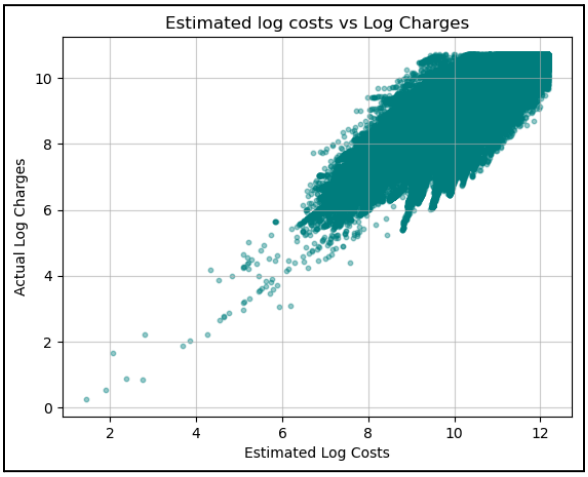

Figure 4.37 Scatter Plot between Log of Estimated cost and Log of Total Charges

As shown in figure 4.36, one can observe a strong positive relationship between the two variables, i.e. as estimated healthcare costs increase, total charges also increase. The coefficient of correlation between the two variables is 0.7753 - Pointing towards a strong positive relation between the 2 variables. Below we also estimate the relationship between logs of the variables to understand whether changing of scale impacts the relationship or not.

From the scatter plot of log of estimated costs and log of Total charges as shown in figure 4.37, we see a strong positive relationship between the two variables, i.e. as log of estimated healthcare costs increase, log of total charges also increase. The coefficient of correlation between the two variables is 0.8236 - Pointing towards a strong positive relation between the 2 variables.

Chapter 5: Results and Discussions

5.1 Introduction

This chapter focuses on the complete implementation and evaluation of the Machine learning models to predict high-cost patients and understanding important features that help in determining this classification. Building on the data exploration and preparations, this chapter outlines the modelling pipeline, starting from transformation of variables, leading to development, tuning, evaluation and interpretation of the machine learning model. This chapter structurally analyzes how the models performed, what are factors driving predictions of high-cost patients.

Section 5.2 discusses the final model selection which is based on observations from exploratory data analysis and the rationale behind the choice of model. Section 5.3 addresses the creation of pertinent derived variable(s), the transformation of dependent variables along with the target variable to fit the model, Section 5.4 details the sampling technique—more especially, the use of stratified train-test split and k-fold cross-validation to guarantee strong model validation. Section 5.5 details the implementation of the XGBoost Classification Algorithm and the training procedure employed.

Core results of the model are present in Section 5.6, which evaluates the base model incorporating all the features. Section 5.7 discusses SHAP to provide a detailed interpretation of feature importance and their influence across predicted cost classes. In order to enhance the model transparency, section 5.8 discusses an alternate model which excludes the total cost feature with the aim to check whether the model can perform reliably on clinical and demographic predictors. Section 5.9 provides a comparative view of both the models in terms of different evaluation criteria.

Finally, Section 5.10 discusses the limitations, assumptions, and model choices made during the study along with considering broader analytical, theoretical and practical implications. This chapter aims at contributing an original and multi-faceted perspective on the constructive role

that machine learning can provide in healthcare cost classification and blending quantitative evaluation with real-world relevance and interpretive depth.

5.2 Model Selection and Rationale

Originally intended to predict healthcare costs, the study took into account both regression and classification models. But it was found during the exploratory data analysis that continuous variables were quite skewed with notable outliers at both the ends. This possesses significant challenges to regression-based modelling where presence of extreme values can distort model behavior, reduce predictive power of the model and increase risk of overfitting.

The presence of outliers were effectively treated by removing the relevant data from the dataset. However, the dataset had a further bigger issue that posed a great challenge to the regression-based modelling. Most of the features present in the dataset were categorical in nature and are rich in clinical and administrative information. However, to align such variables with regression-modelling extensive transformation or encoding of variables is required. One-hot coding of such variables can result in a high-dimensional sparse matrix, which can malign the performance and interpretability of the regression model.

To address this challenge and to better align with the objectives of the research, the problem was reframed as a multi-class classification problem from a regression problem. The target variable, Total Charges was accordingly treated to suit the model requirements. This approach provides an optimal balance between class separability, interpretability, and model stability.

The classification model chosen for the research was XGBoost Classification, a gradient-boosted tree ensemble known for its robustness, scalability, and strong performance on tabular data, especially with high-cardinality categorical features, missing values, and the need for strong predictive performance.

The focus of the subsequent sections is to implement a classification framework, evaluate its predictive power and interpret the outcomes using relevant metrics and visualisations.

5.3 Variable Transformation and Feature Engineering

The predictive power of any machine learning model depends heavily on quality and structure of the input data. Therefore, a concentrated effort is required in transforming both target and independent variables.

5.3.1 Transformation of Target Variable

The target variable, total charges, originally a continuous numeric variable, was transformed using tertile-based binning to convert it into a categorical classification target. This transformation allowed for a more stable modeling framework by eliminating the influence of outliers on the model performance along with facilitating clearer segmentation of patients into low, medium and high cost groups.

5.3.2 Creation of Derived variables

In order to capture more nuanced patterns in patient utilisations and clinical complexities, variables derived from existing ones were created. One such variable is Cost per day, obtained by dividing total costs by (length of stay + 1) to normalize cost with respect to hospitalization duration and avoid division by zero. Additionally, a log transformation of Total Costs and Total Charges were initially explored to reduce skewness, although it was ultimately used only for exploratory visualization and not retained in the final model..

5.3.3 One-hot labelling

Categorical variables, which consist a substantial chunk of the dataset, were converted into machine-readable format using one-hot labelling. This procedure generated a binary column for every feature category, so augmenting the dataset's total number of predictors. This produced a high-dimensional feature space but let the model fully exploit the predictive ability of categorical distinctions.

Following all the above mentioned transformations and feature engineering, the resulting dataset consisted in 11,99,957 rows and 111 independent features. Out of 111 features, 3 continuous variables and 108 engineered and encoded features, covering demographic, clinical and service-related aspects of the patient data. These features lay a foundation for creation of the classification model and play a pivotal role in further model training, evaluation and interpretation

5.4 Sampling and Validation Strategy

Ensuring that machine learning models generalise effectively to unprocessed data depends on a strong sampling and validation technique. For the study, a stratified train-test split was used whereby 80% of the data was used for training and the other 20% for testing. Using stratification based on the target variable, all three low-cost, medium-cost, and high-cost classes—that is, proportionately represented in both subsets—are guaranteed. This strategy avoided class imbalance from biassing outcomes or distorting the training process.

5-fold-cross-validation on the training data was carried out in order to lower overfitting risk and so strengthen the model's dependability. Training data was split in this method into five equal sections; the model was trained on four of them and validated on the remaining part in every iteration. Five times this procedure was carried out, once using each of the five subsets as a validation fold. The average performance over the five folds gave a more consistent and generalised estimate of the predictive capacity of the model.

SHAP values were computed on the whole 2.4 lakh row test dataset for the aim of model interpretation. Unlike many earlier studies using subsets of the test set for SHAP computation, this work used the whole test dataset to guarantee complete and objective interpretability. This enhanced the quality of visualisations and insights gained from SHAP analysis as well as enabled an accurate evaluation of feature importance over all predicted classes.

Stratified train-test splitting, cross-valuation, and full test set-based SHAP computation taken together guaranteed statistical rigour and interpretive depth in the evaluation of the classification model. These validation methods help to support the generalisability and credibility of the results expressed in the later sections.

5.5 Model Implementation

The final classification model was implemented using XGBClassifier from XGBoost library. The model was configured to predict patient membership in one of the three cost categories - Low-cost category (Class 0), mid-cost category (Class 1) and High-cost category (Class 2) - based on the tertile transformation of the original variable. The input data consists of 111 features out of which 3 are numeric and 108 binary categorical features.

The model was initialised using default parameter settings, with no manual tuning applied. Despite the lack of hyperparameter optimisation, the model delivered a strong and consistent performance across test data and cross-validation. Owing to its reliability and practical accuracy, this model configuration was adopted as the final deployed model for the classification task.

Once the model was deployed, its evaluation was carried out using standard classification metrics—accuracy, precision, recall, F1-score, and AUC-ROC—to comprehensively assess performance. These results are presented and analyzed in the next section.

5.6 Evaluation of Base Model

The first version of the classification model was trained with all 111 features which includes Total Cost as a feature. This model was evaluated on a hold-out test dataset using multiple performance metrics and supported by 5-fold cross-validation to assess generalizability. The results indicate that the XGBoost classifier delivered strong predictive performance across all classes.

The model achieved a strong accuracy of 81.67% with Macro-averaged precision, recall and F1-score values of 81.79%, 81.66% and 81.70% respectively. The AUC-ROC score was 0.9456, which showcases the high discriminative ability across the three cost classes of the classification model.The results of model evaluation are summarised in the Table 5.6.1 given below

Table 5.1 Summary of Evaluation Metrics

Metric	Value
Accuracy	81.67%
Precision (Macro)	81.79%
Recall (Macro)	81.66%
F1-Score (Macro)	81.70%
AUC-ROC (OvR Macro)	0.9456
Cross-Val Accuracy	81.56% ± 0.0006

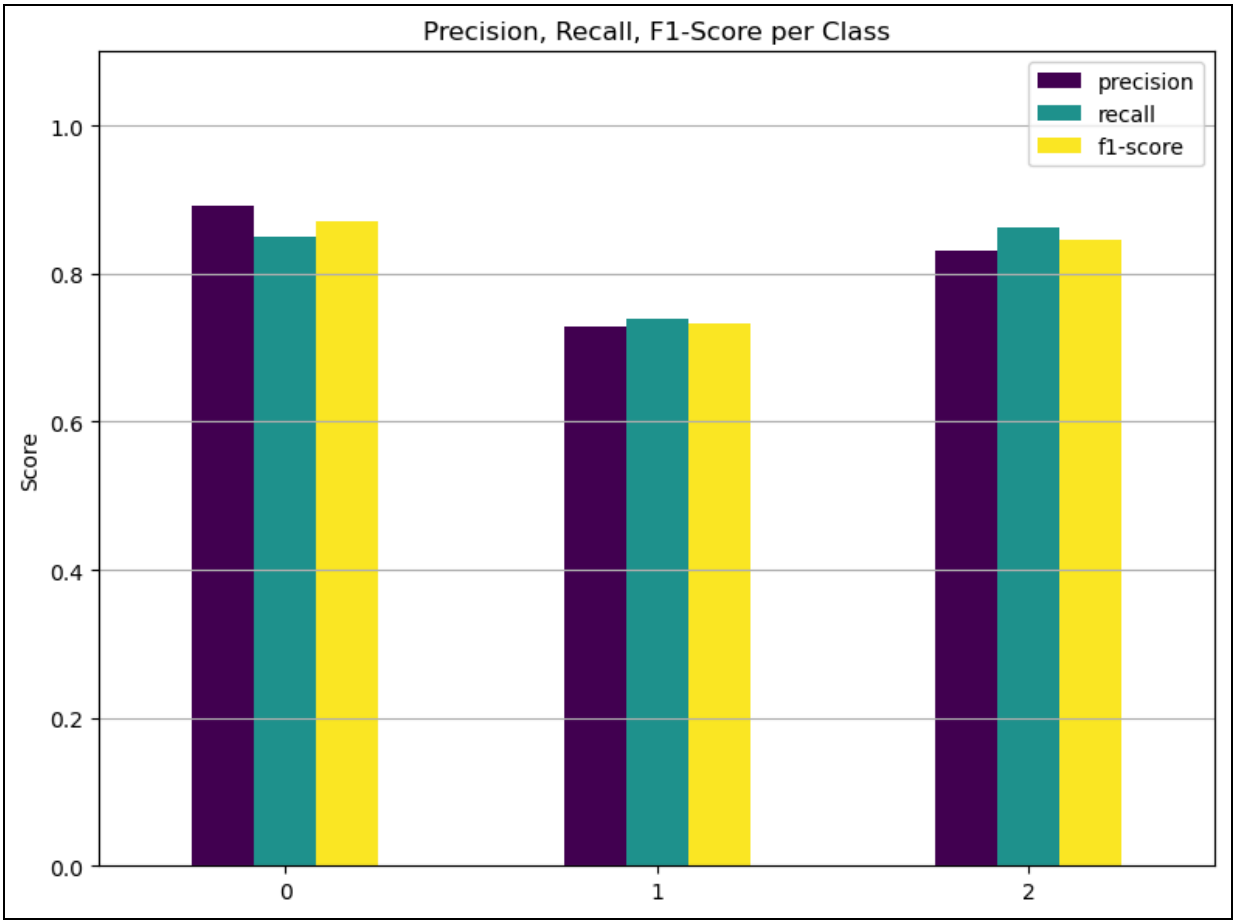

Figure 5.1 Class-wise Evaluation Metrics

To gain a deeper understanding of per-class behaviour, a bar plot of the classification report was generated, visualising and summarising precision, recall and f1-score for each class (low-cost, medium-cost and high-cost). As shown in figure 5.6.2 the model performed the best on low-cost class (class 0) followed by high-cost class (class 2) with all the three evaluation metrics score above 0.8. However, the scores for medium-cost class (class 1) were slightly lower compared to the other two classes. This can be explained by the fact that mid-cost patients have overlapping feature characteristics with other two classes.

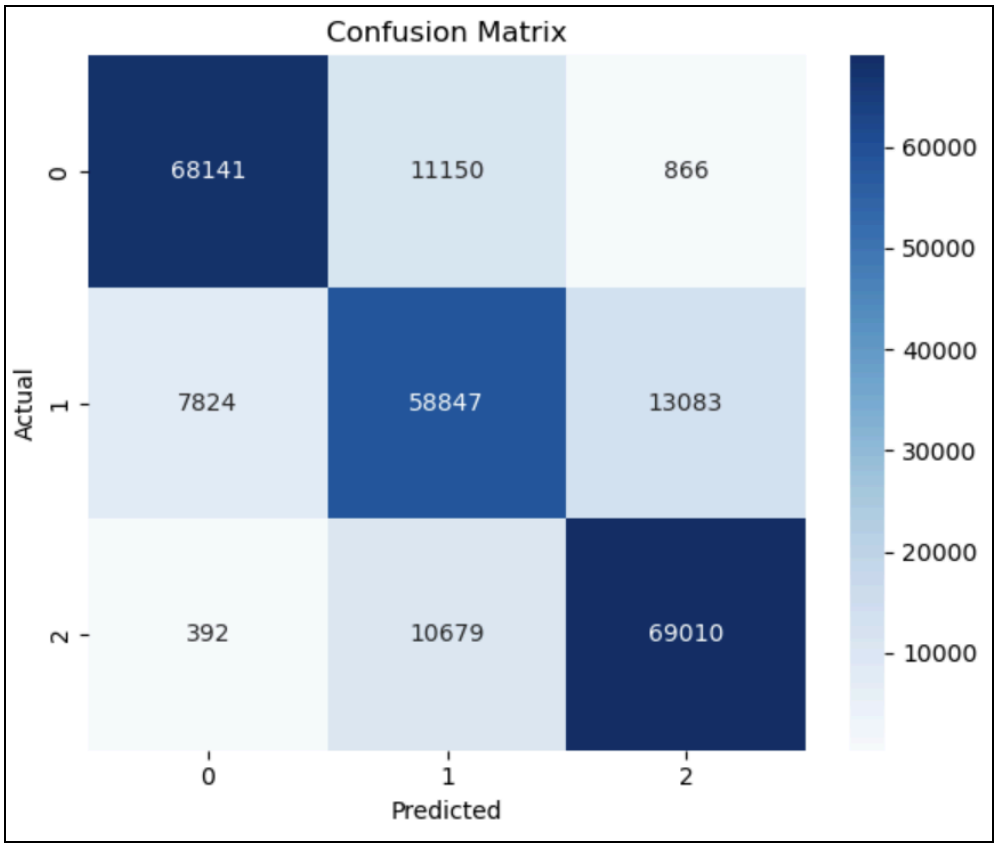

Figure 5.2 Confusion Matrix

The confusion matrix, as shown in figure 5.6.3 further reinforces the observation made above, displaying the raw count of predictions for each actual vs. predicted class. In this, one could easily observe that most of the mis-classification has been w.r.t to the class 1, i.e. mid-cost patients with 17.81%. This is a common feature in cost categorisation where clinical and financial boundaries can be ambiguous.

To further assess how well the model performs across all the thresholds, ROC curves were plotted for each class in a one-vs-rest setup (Figure 5.6.4). Each curve showed high AUC values, reinforcing the model's ability to distinguish among all three classes with confidence.

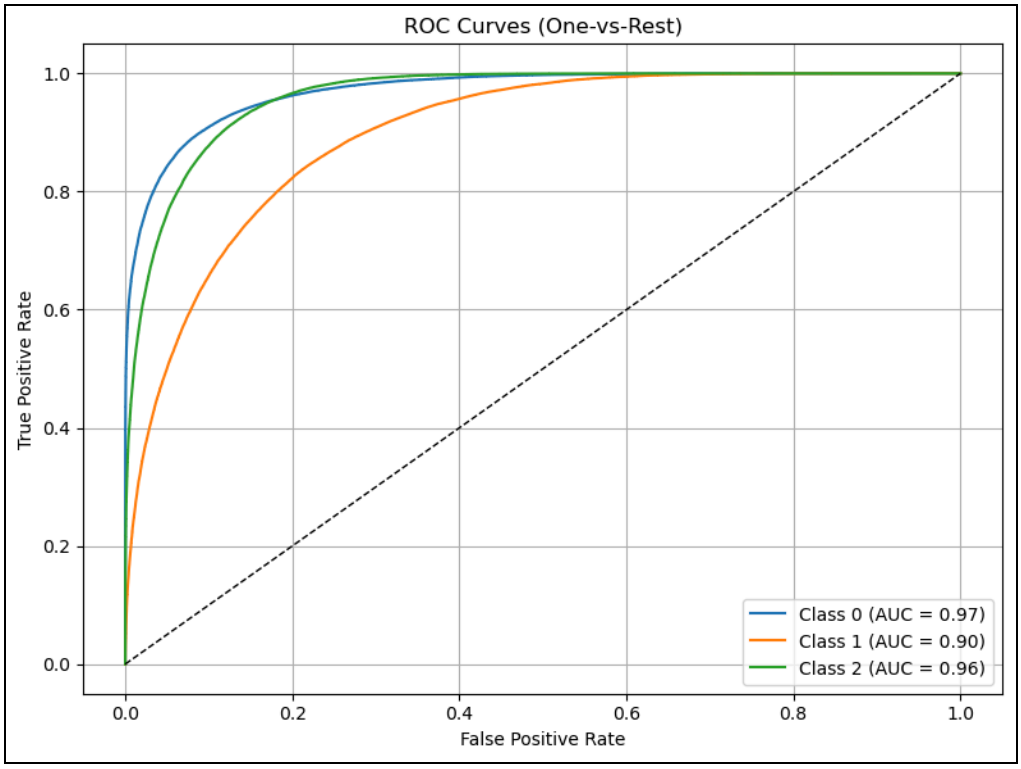

Figure 5.3 ROC Curves for each class

Model consistency was further validated using 5-fold cross-validation, with results showing a mean accuracy of 81.56% ± 0.0006. A visual summary of fold-wise performance is provided in Figure 5.6.5, highlighting the stability and generalizability of the model.

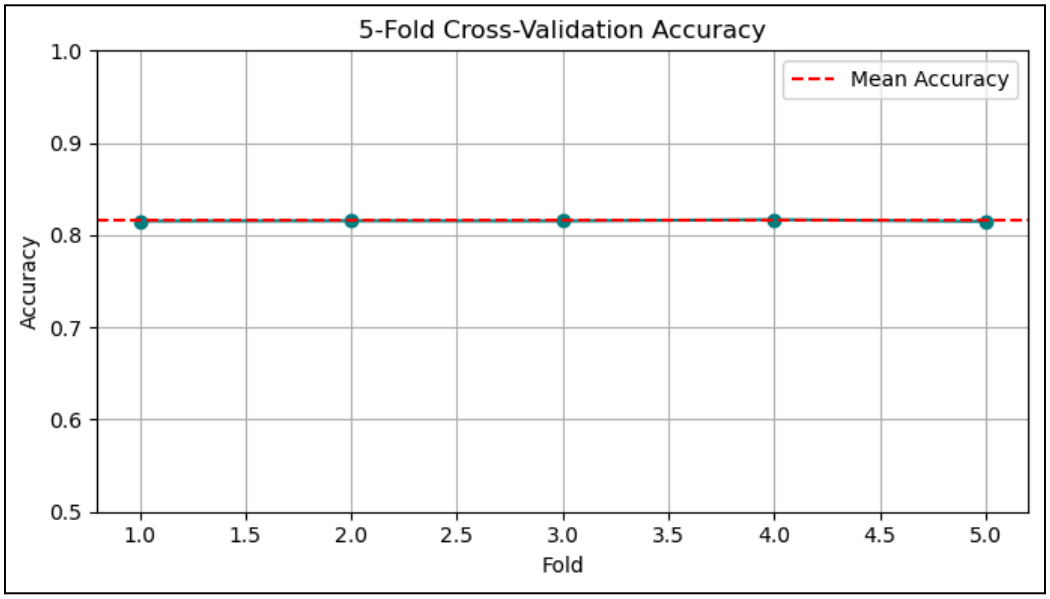

Figure 5.4 5-Fold Cross-Validation Accuracy per Fold

Model consistency was further validated using 5-fold cross-validation, with results showing a mean accuracy of 81.56% ± 0.0006. A visual summary of fold-wise performance is provided in Figure 5.6.5, highlighting the stability and generalizability of the model.

5.7 SHAP-Based Interpretability of Model

Model interpretability plays a vital role in healthcare-related machine learning models, especially when predictions influence financial and clinical decisions. To explain how XGBoost classifier made certain predictions in the model, SHAP (SHapley Additive exPlanations) was employed. SHAP allows for a unified measure of feature importance based on game theory, offering insights into both global feature impact and class-specific contributions

The SHAP values were computed on the entire test dataset to ensure consistency and generalizability of interpretations. The resulting stacked SHAP summary plot, presented in Figure 5.7.1, displays the top features driving predictions toward each of the three cost classes (low, mid, and high), along with their relative magnitude and distribution shown in Figure 5.7.2, Figure 5.7.3 and Figure 5.7.4.

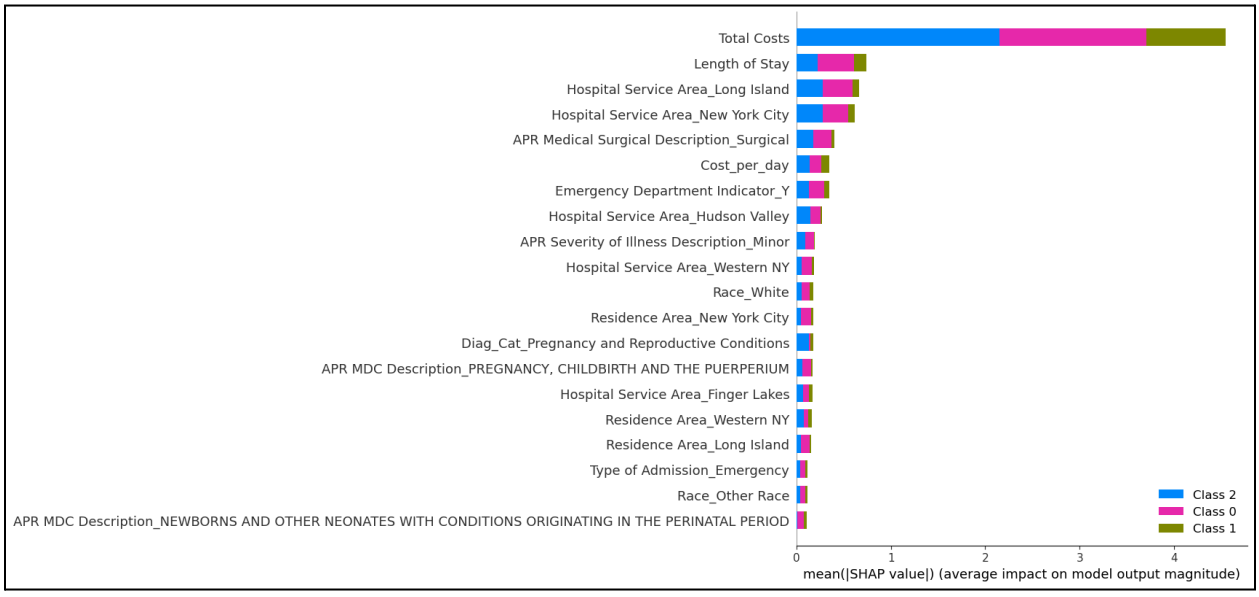

Figure 5.5 SHAP Summary Plot – Overall Feature Impact

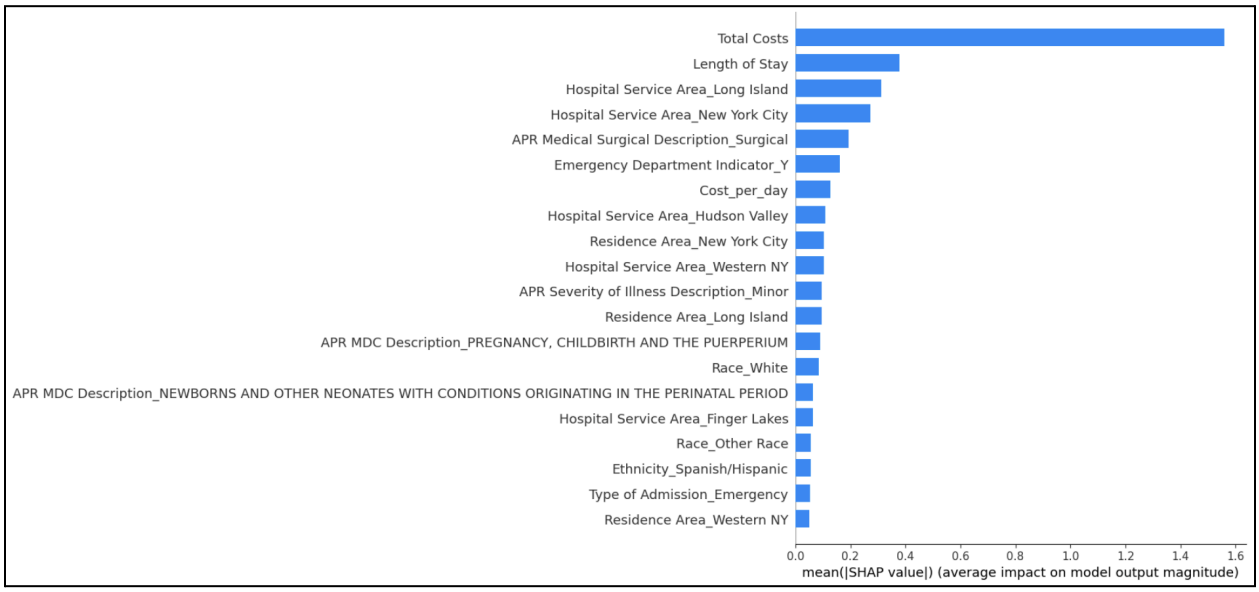

Figure 5.6 SHAP Summary plot for Class 0 (low-cost)

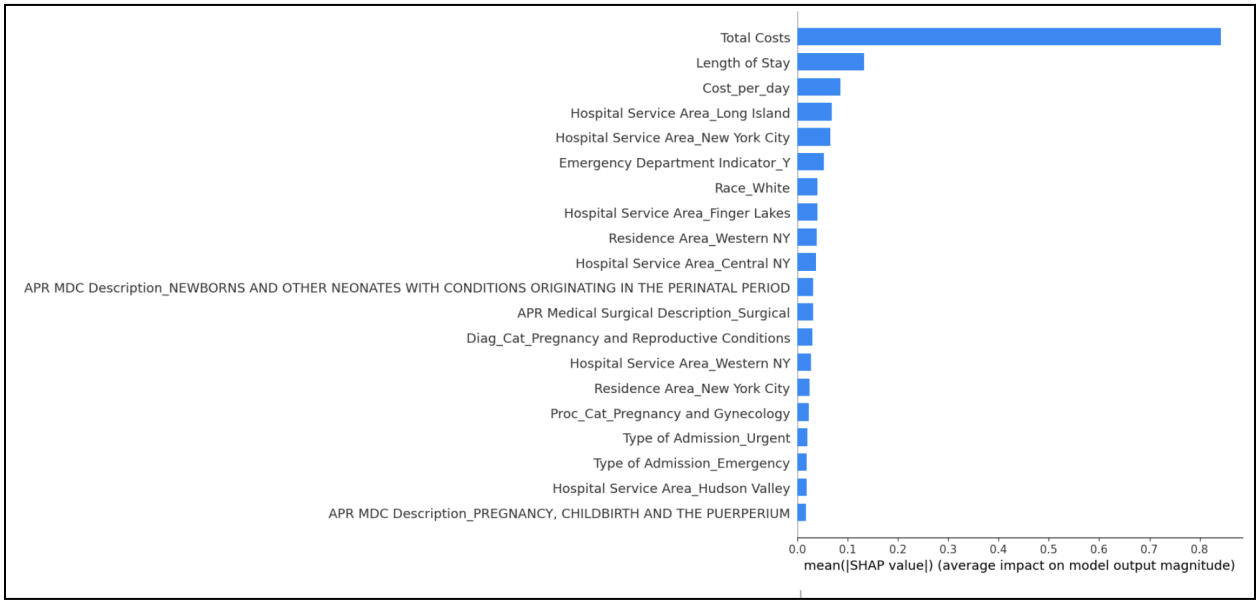

Figure 5.7 SHAP Summary plot for Class 1 (medium-cost)

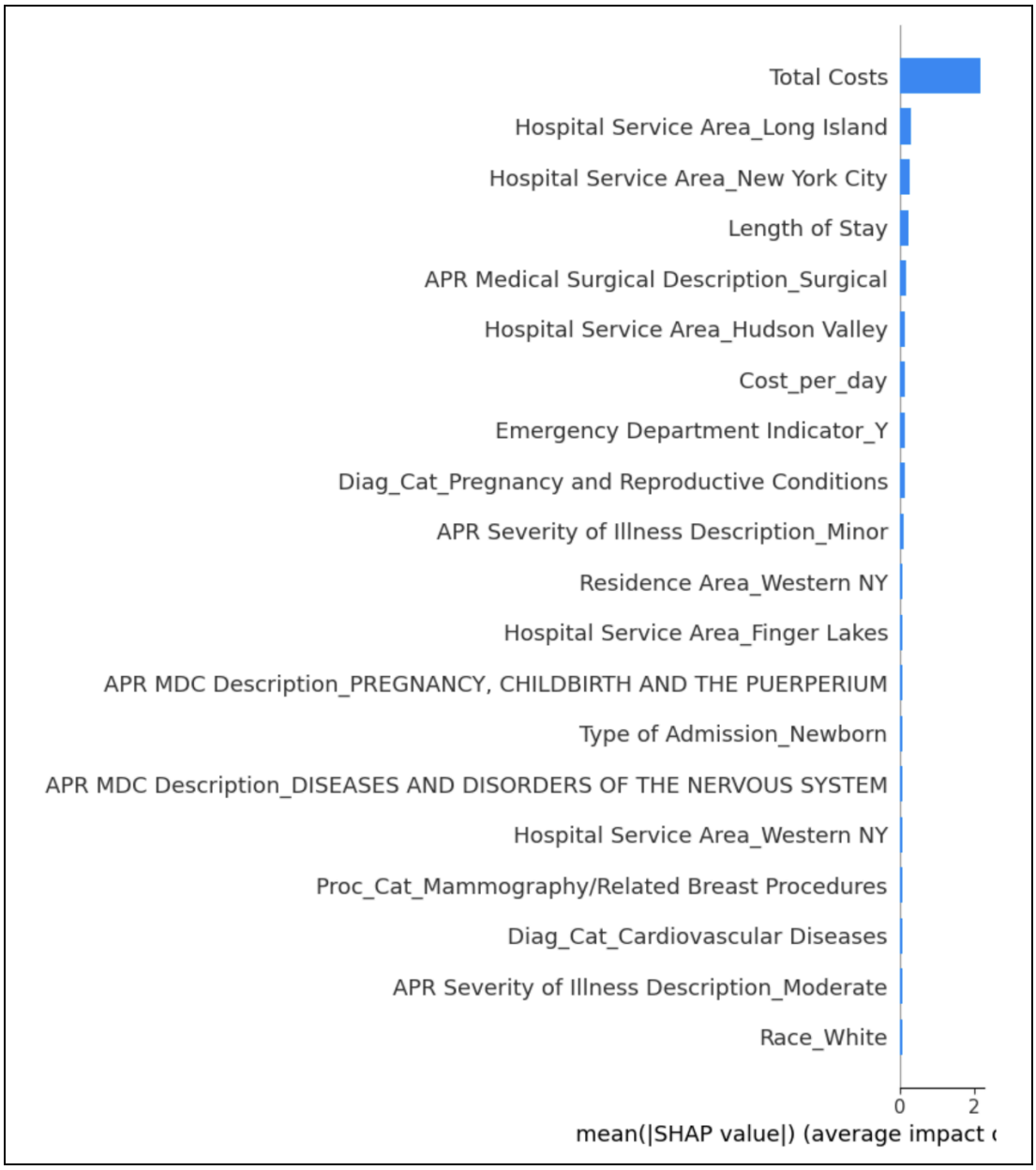

Figure 5.8 SHAP Summary plot for Class 2 (high-cost)

Aligning with the pre-modelling expectations, total costs is a dominant feature that strongly influences the decision making of the model. However, the feature had the highest share of influence in high-cost predictions (Class 2), reinforcing the concern of possible target leakage. While its inclusion improved accuracy, it also risks obscuring the contributions of more independently informative variables.

Other than total costs, the model consistently relies on features like length of stay, cost per day, Hospital Service Area and APR Medical Surgical Description,each showing varying degrees of influence across the classes. For example:

- Length of Stay and cost per day showed relatively balanced influence across all classes, reflecting their strong explanatory power for healthcare expenditures.
- Geographic features like Hospital Service Area and Residence Area had stronger influence in Classes 1 and 2, suggesting regional disparities in healthcare delivery and associated costs.
- Clinical descriptors, such as Emergency Department Indicator and APR Severity of Illness, contributed notably to mid- and high-cost classifications, aligning with expectations that emergency and complex cases tend to incur higher costs.

The stacked bar representation of SHAP values provides a more comparative and holistic view than class-wise plots, enabling simultaneous assessment of both feature magnitude and class-specific direction. This visualization supports the conclusion that, although performance was high, the model's interpretability is heavily skewed by one dominant variable.

This analysis underscores the need to evaluate an alternative model that removes Total Costs to surface other latent patterns and assess model robustness without reliance on a target-derived feature. The performance and interpretability of that model are discussed in the next section.

5.8 Evaluation of Alternate Model

Given the dominance of total costs in the model, a second XGBoost Classification model was developed excluding this feature to evaluate the **t**rue predictive power of independent demographic, geographic, and clinical variables of the dataset on classifying patients. The objective of developing this alternate model was two-fold; one, to eliminate the reliance of model on a variable that is highly correlated to the target variable and two, to assess whether a high-performing and interpretable model can be developed to classify patients based on features that are more reflective of the real-time patient data.

The alternate model was trained on the same preprocessing pipeline and hyperparameters as the base model to ensure comparability between the two models and compare their metrics. Key

features like Length of Stay and cost per day were retained as they were moderately related to the target variable (respective coefficients of correlations were 0.45 and 0.49) along with the categorical variables.

On running the model on the test dataset, it achieved a strong accuracy of 81.27% with Macro-averaged precision, recall and F1-score values of 81.43%, 81.26% and 81.32% respectively. The AUC-ROC score was 0.9438, slightly lower than the previous model.The results of model evaluation are summarised in the Table 5.8.1 given below. Overall, the results justify the observation that exclusion of total estimated costs from the dataset did not led to no significant loss in overall predictive performance, affirming that other features collectively hold meaningful signals in determining patient cost classification.

Table 5.2 Summary of Evaluation Metrics for Alternate model

Metric	Value
Accuracy	81.27%
Precision (Macro)	81.43%
Recall (Macro)	81.26%
F1-Score (Macro)	81.32%
AUC-ROC (OvR Macro)	0.9438
Cross-Val Accuracy	81.27% ± 0.0009

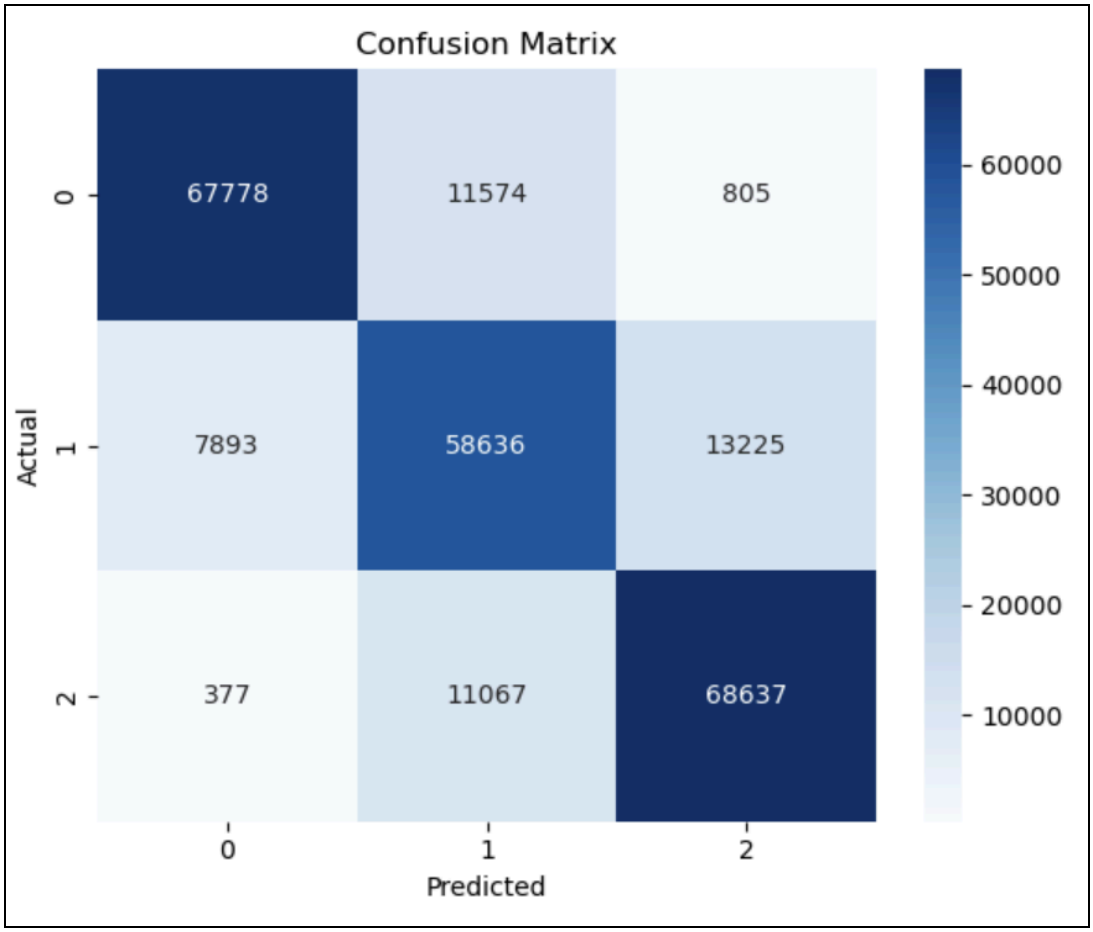

Figure 5.9 Confusion Matrix of alternate model

The confusion matrix of the alternate model, as shown in figure 5.9 above points to the fact that the mis-classification pattern of both the models is the same. In the alternate model also, most of the mis-classifications are concerned with the class 1, i.e. the medium-cost group. This further solidifies the observation that classification between medium and other cost groups can be ambiguous given clinical and demographic data.

Figure 5.10 Class-wise Evaluation Metrics of alternate model

Figure 5.10 shows the class-wise metrics of the alternate model. From the graph, it can be inferred that the metrics of both the models are on the similar level with low-cost and high-cost classes having values of metrics above 0.8 and medium-cost class performing slightly less than the other two classes.

Figure 5.11 given below plots the ROC curves of each class were plotted for the alternate model. The AUC-ROC score for each class is still high and similar to those in the previous model.

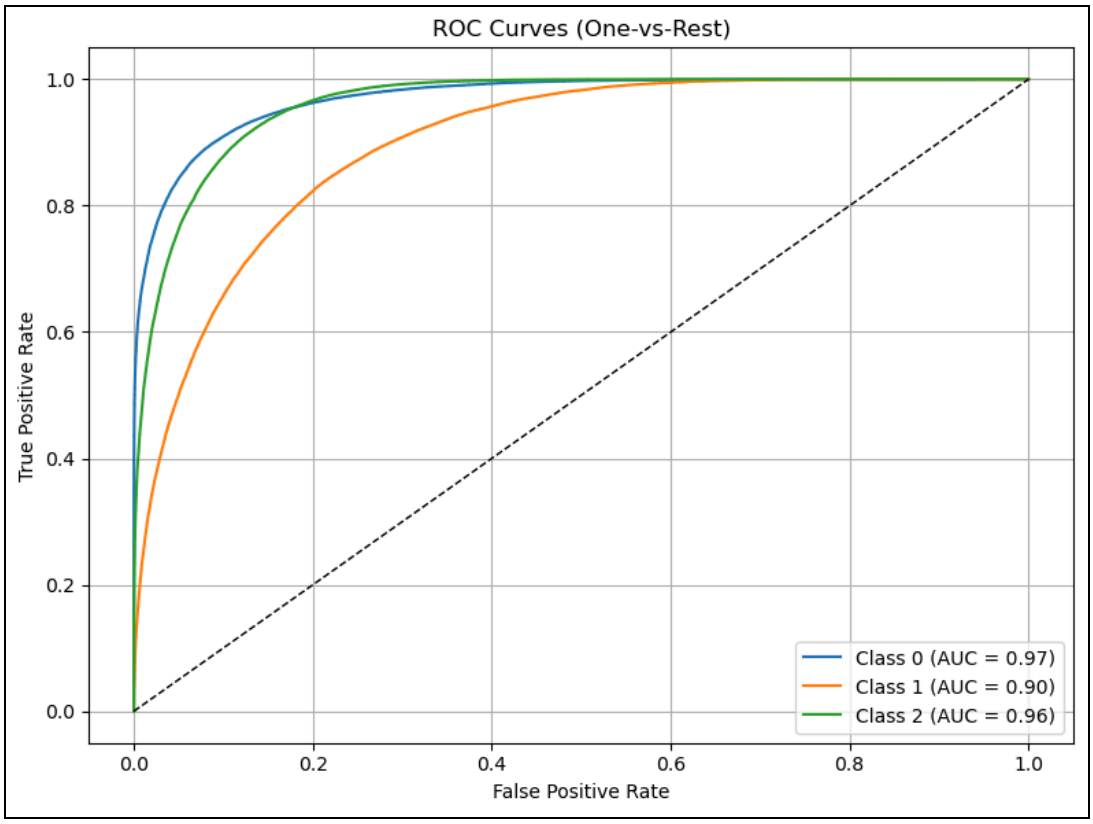

Figure 5.11 ROC Curves for each class in alternate model

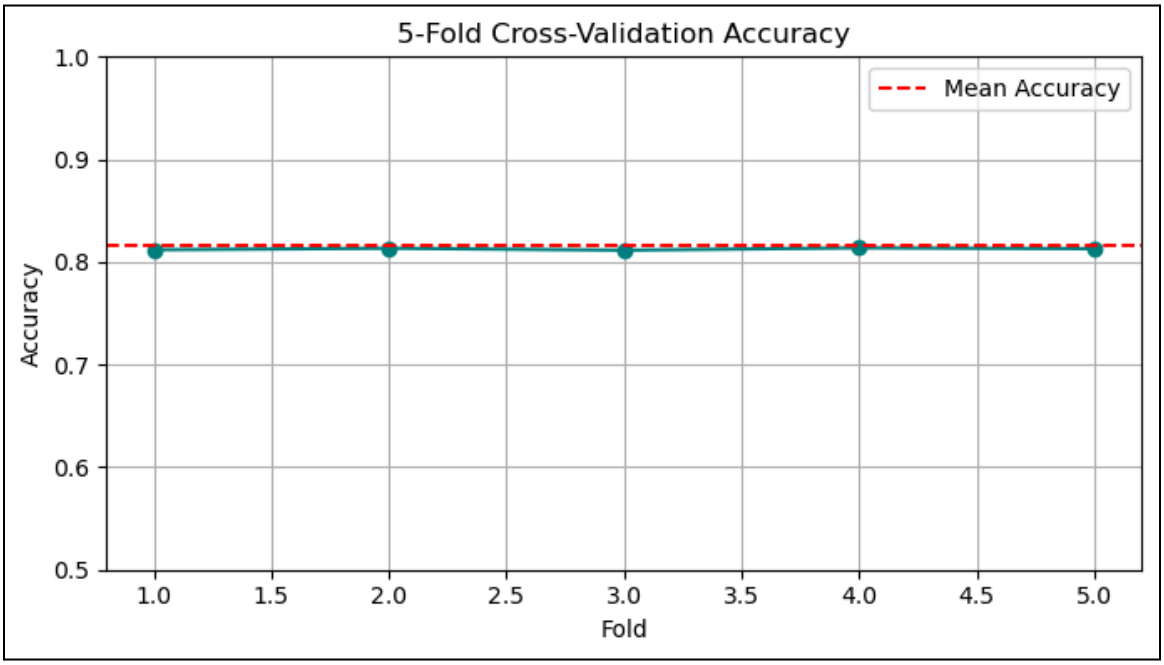

Figure 5.12 5-Fold Cross-Validation Accuracy per fold for alternate model

Further validation using 5-fold cross-validation confirmed the stability of this model, with a mean accuracy of 81.27% ± 0.0009, as shown in Figure 5.12. The low variance in fold-wise performance reinforces the generalizability of the results even without access to Total Costs.

To explain how the alternate model achieved such strong performance despite the exclusion of Total Costs, SHAP values were computed over the entire dataset as shown is figures 5.13 to 5.16 reveals a significant shift in the feature reliance of the model:

- Cost per day and Length of Stay emerged as the two most influential features, effectively replacing Total Costs in driving model predictions across all three classes.
- Other high-impact features included Hospital Service Area, APR Medical Surgical Description, Emergency Department Indicator, and Severity of Illness, demonstrating the model's ability to leverage clinical and geographic inputs.
- The stacked SHAP plot (Figure 5.18) confirmed consistent contribution of these features across classes, with Class 2 (high-cost) showing dominant dependence on high per-day costs and long stays.

This interpretability confirms that even in the absence of Total_Costs, the model not only preserved its predictive power but also surfaced more ethically and operationally actionable cost drivers. It allowed the identification of high-cost risk based on features available at or near the time of admission—an essential trait for early intervention systems.

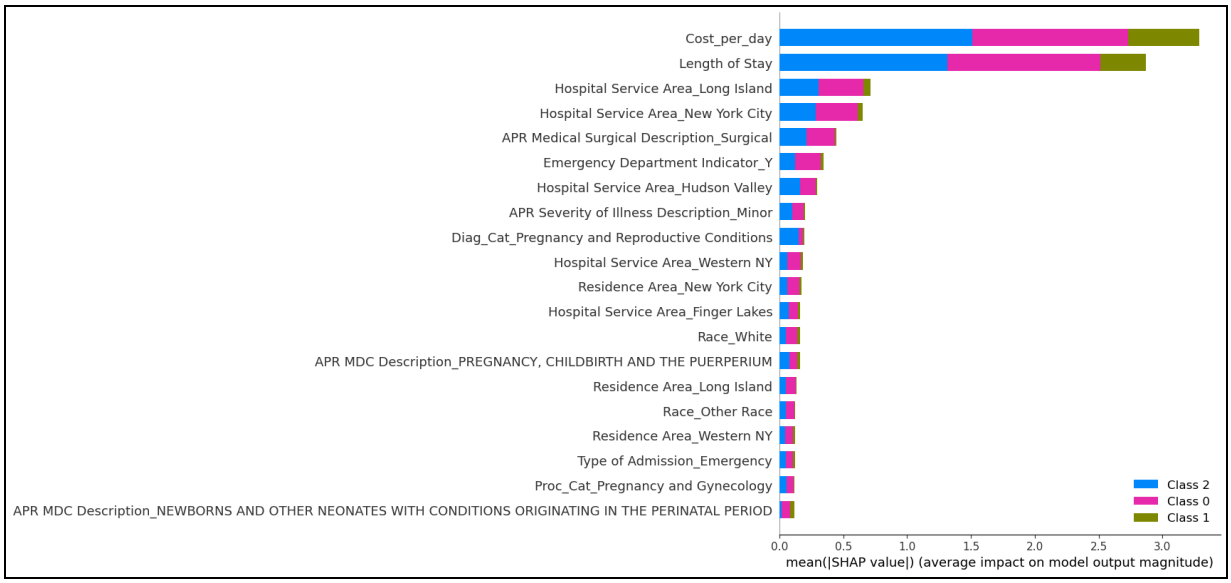

Figure 5.13 SHAP Summary Plot – Overall Feature Impact

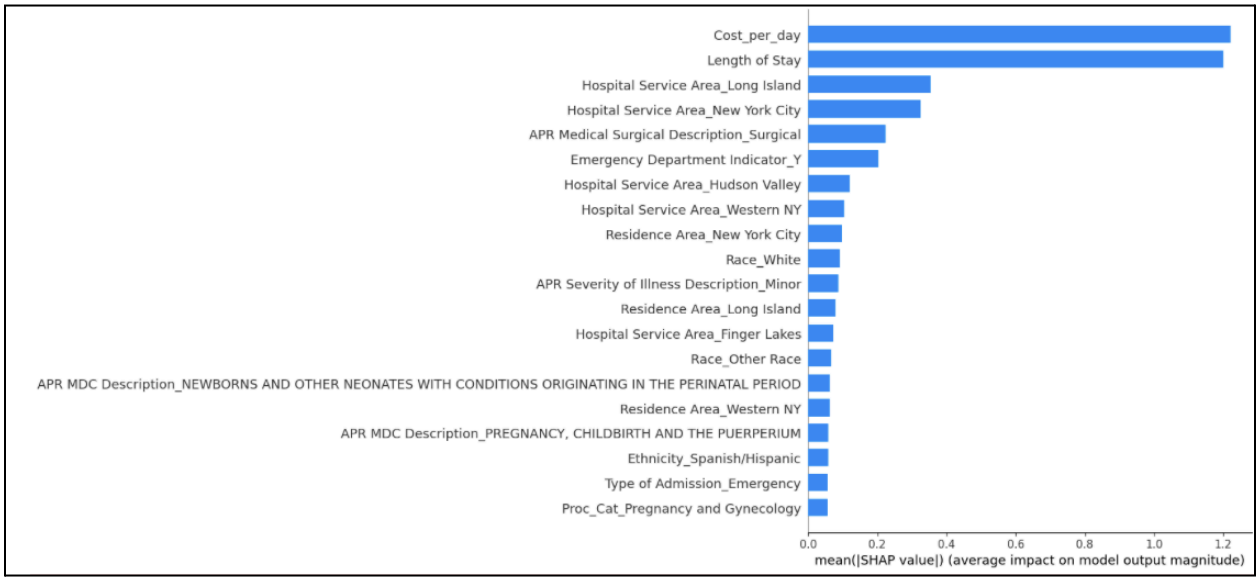

Figure 5.14 SHAP Summary plot for Class 0 (low-cost) Alternate model

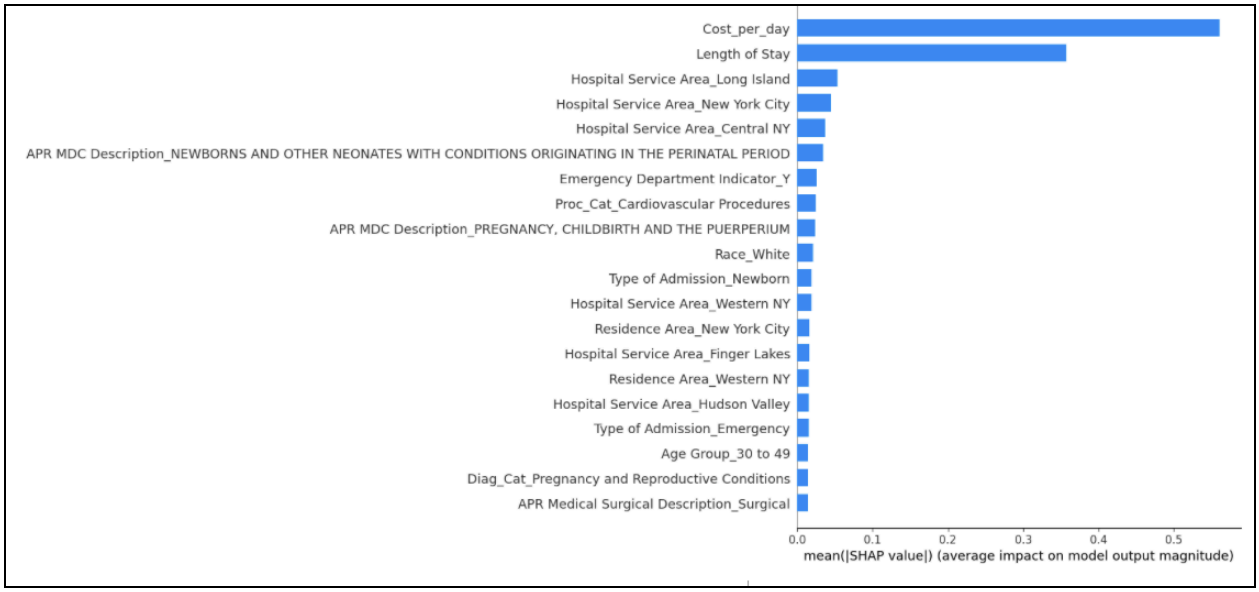

Figure 5.15 SHAP Summary plot for Class 1 (Medium-cost) Alternate model

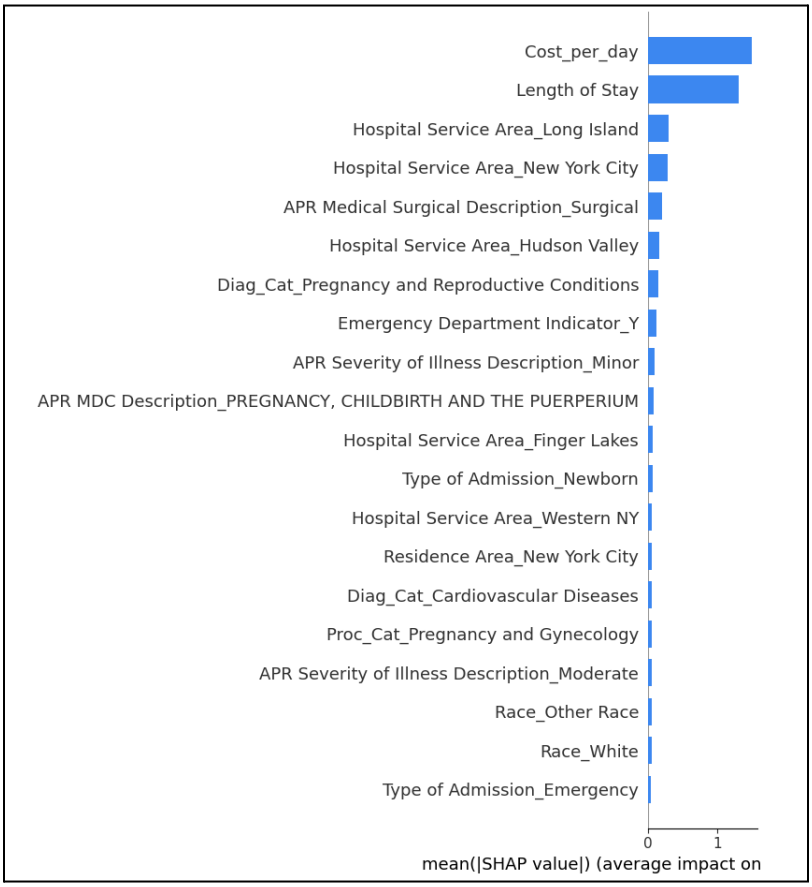

Figure 5.16 SHAP Summary plot for Class 2 (High-cost) Alternate model

5.9 Comparative Analysis of Two models

This section provides a detailed comparison of the two XGBoost Classifications models build during the research: first, the base model which includes Total cost as a feature and second, the alternate model, which excludes the total cost feature in order to enhance the interpretability and applicability of the model. The comparison is based on predictive performance metrics and explainability insights from SHAP analysis.

5.9.1 Performance Metrics Comparison

On the overall level, both the models demonstrated strong predictive power with comparable evaluation metrics.The base model achieved an overall accuracy of 81.67% , F1-score of 81.70% and AUC-ROC score of 0.9456. On the other hand, the alternate model had an overall accuracy of 81.27%, F1-score of 81.32%, and AUC-ROC of 0.9438. The summary is given below in table 5.3.

Table 5.3 Comparison Table

Metric	Base Model	Alternate Model
Accuracy	81.67%	81.27%
Precision	81.79%	81.43%
Recall	81.66%	81.26%
F1-Score	81.70%	81.32%
AUC-ROC	0.9456	0.9438
CV Accuracy	81.56% ± 0.0006	81.27% ± 0.0009

While the performance of Alternate model is slightly less than that of Base model, the differences can be tolerated by the interpretability advantage gained in the alternate model.

5.9.2 SHAP Analysis and Model Interpretability

In the base model, SHAP analysis revealed that Total Estimated Costs overwhelmingly dominated the prediction of the 3 classes. While translated into slightly better performance metrics, it risked information leakage due to its high correlation to the target variable. As a result, the model's decisions were less clinically actionable, especially for early-stage intervention.

In contrast, the alternate model redistributed importance across multiple features. SHAP visualizations showed that:

- Cost per day and Length of Stay became the most influential cost proxies,
- Clinical variables like APR Surgical Description, Emergency Indicator, and Severity of Illness played stronger roles, and
- The model utilized diverse inputs rather than over-relying on a single financial metric.

This shift improved the model's transparency, trustworthiness, and ethical robustness, making it more appropriate for real-time healthcare decision-making where Total Costs would not yet be known.

5.9.3 Practical Relevance and Final Verdict

From the applied perspective, the alternate model provides greater practical utility in hospital operations, insurance underwriting, and patient triage systems. It allows decision-makers to act early, based on available and interpretable features, rather than retrospective billing data.

Thus while the base model outperforms the alternate model on paper, the alternate model avoids information leakage, Offers superior explainability, aligns with real-world clinical workflows, and enables earlier risk detection.

Accordingly, the alternate model is recommended as the more robust and deployable solution for identifying high-cost patients using structured inpatient data.

5.10 Limitations, Assumptions, and Broader Implications

While the research successfully demonstrates the use of machine learning, especially XGBoost Classification to predict high-cost patients using structured inpatient data, several limitations and assumptions need to be acknowledged to frame the findings within appropriate bounds. This section explores the broader implications for healthcare research and real-world implementation.

5.10.1 Assumptions and Modelling Decisions

The study adopted several assumptions during model designing and implementation. First, it was assumed that the historical inpatient discharge data can serve as a stable and reliable proxy of the future cost behaviour and patients classification. All the models were trained and tested on this dataset with the assumption that the features available at the discharge are sufficient for classifying patients according to the costs incurred.

Second, to isolate the overwhelming influence of total costs on the modelling decisions, an alternate model was built excluding this variable. This was based on the critical assumption that proxy variables like cost per day and length of stay and other categorical factors could capture and understand the underlying cost behaviour without any information leakage. This decision enabled a trade-off between accuracy and interpretability, ensuring that the model remains viable for prospective,early-stage prediction tasks in clinical or administrative settings.

Third, the categorical variables of the dataset were transformed using one-hot coding, resulting in a large feature space. While this allowed for algorithmic compatibility, it also introduced potential issues of multicollinearity and dimensional sparsity, which could compromise the robustness of the model. However, XGBoost's tree-based architecture naturally mitigates many of these concerns by handing correlated and high-cardinality features.

5.10.2 Limitations of the Study

Despite the robustness of the XGBoost Classifier, this study is constrained by a few limitations:

- **Generalizability:** Model was trained on a dataset derived from one state's healthcare system (New York), and its performance might not extrapolate well to other states or countries with different healthcare systems, patient demographics or billing systems.
- **Feature Granularity:** Many features of the model were derived from categorical tags. These may abstract awat clinical nuances that could further improve prediction.
- **Temporal Dynamics Ignored:** The analysis does not incorporate time-series data or trends in patient health over time, which have the potential to add to the predictive power for cost forecasting.
- **Interpretation Challenge:** While SHAP improves transparency, the use of dummies and complex engineered features still limits the layperson's ability to interpret or act on individual predictions in a clinical workflow

5.10.3 Analytical and Theoretical Contributions

From a theoretical standpoint, thai research shows that cost prediction can be effectively framed as a classification task and not just a pure regression problem. This reframing simplifies the understanding of model and aligns better with the operational decision-making in hospitals, which often relies on risk groups rather than exact cost values.

Moreover, the research contributes to the growing body of evidence that structured administrative health data-when appropriately pre-processed-can reveal strong signals for cost-based patient segmentations. This supports theories in health economics which asserts that the hospital cost drivers are hidden in demographics, procedural and institutional patterns that than purely financial inputs.

5.10.4 Practical Implications

From a practical viewpoint, the alternate model has clear applicability in:

- **Early hospital triage:** Identifying high-cost patients at or near admissions, enabling better resources allocations and care coordination.
- **Insurance risk stratification:** Supporting insurance underwriters in evaluating patient risk using structured medical histories.

- **Policy Planning:** Informing government and private healthcare bodies about the systemic factors that influence high-cost outcomes.

Most importantly, the ability to develop high-performing model without relying on Total estimated costs showcases the feasibility of deploying understandable and proactive ML systems in healthcare operations.

5.11 Summary

This chapter demonstrates the effectiveness of XGBoost Classification models in predicting healthcare cost categories. While the base model performed slightly better over the alternate, but it reliance on total estimated costs limits the interpretability of other variables. The alternate model, which excludes this feature, maintained comparable performance while offering stronger explainability and practical applicability.

SHAP analysis confirmed that clinically relevant variables like Length of Stay, cost per day and severity indicators played key roles in both the models, especially in absence of direct cost data.

Overall, the alternate model emerged as the more balanced solution-accurate, understandable and deployment-ready.

Chapter 6: Conclusions and Future works

6.1 Introduction

This chapter offers a reflective summary of the entire study, placing its outcomes with broader academic and practical context. Moving ahead from model specific evaluations, it highlights how the research contributes to various domains such as healthcare delivery, data science, academic inquiry, and policy development. It also identifies areas where future research and implementation can build on the foundation laid in this work.

6.2 Discussion and Conclusion

This study set out to explore whether the machine-learning models could reliably predict high-cost patients using structured inpatient data. Through a streamlined process of problem framing, feature engineering, model development and evaluation, it was demonstrated that it is possible to classify patients into cost-risk categories even in absence of actual cost values.

By developing both a performance-optimised based model and a more interpretable alternate model, the research stressed on the need to balance accuracy with interpretability. The alternate model, while slightly less precise, proved to be a more ethical and deployable solution-relying solely on features available at point of care.

Overall, the study affirms that cost-based classification is a viable strategy for operations planning, financial forecasting and risk management in the healthcare sector. It also reinforces the strong need for interpretability when machine-learning methods are applied to real-world decision making.

6.3 Contribution to Knowledge

The value of this research lies in its multidimensional impact across disciplines and stakeholder groups:

6.3.1 Academics and Research

- Introduced a new perspective by reframing healthcare cost prediction as a classification problem.
- Validated the effectiveness of models that exclude target-adjacent features, contributing to literature on ethical AI design.
- Demonstrated the use of SHAP as a tool for powerful model interpretability, enabling transparent exploration of healthcare cost drivers.

6.3.2 Healthcare Providers and Hospital Administrators

- Showed that hospitals can identify high-cost patients using admission-time features like Length of Stay, cost per day, severity and risk descriptors.
- Offers a ready-to-integrate model structure that aligns with operational workflows, triage decisions, and resource allocation strategies.

6.3.3 Data Scientists and ML Engineers

- Provided a scalable, modular framework for pre-processing, training, validating, and interpreting ML models on structured hospital data.
- Demonstrated how proxy features can replace sensitive or unavailable variables without compromising predictive power of the model.

6.3.4 Policymakers and Public Health Authorities

- Contributed an ethical, explainable framework for early risk prediction—useful in designing cost containment programs or prioritizing preventive care.
- Prepares the case for using interpretable models in resource-limited settings where real-time cost data may not be accessible.

6.4 Future Recommendations

In order to extend the relevance and application of the research in real world, following recommendations are proposed according to different sectors:

6.4.1 Healthcare Systems and Hospitals

- **EHR Integration:** Inclusion of the model in the hospital information systems can trigger real-time alerts for high-risk patients.
- **Clinical Decision Support**: Pair model outputs with intervention protocols such as early discharge planning or social work consults.

6.4.2 Insurers and Risk Managers

- **Utilize for Claims Triage**: Use cost classification predictions during pre-authorization to prioritize high-risk cases.
- **Cost Control Programs**: Leverage model insights to design preventive outreach or bundled payment strategies.

6.4.3 Data Science Teams

- **Explore Temporal Models**: Incorporate patient visit histories and longitudinal patterns for enhanced accuracy.
- **Fairness Audits**: Test for performance disparities across race, gender, insurance status, and correct for any systemic bias.

6.4.4 Academia and Educators

- **Curriculum Inclusion:** Integrate this work as a teaching case in healthcare analytics and applied machine learning courses.
- **Interdisciplinary Research:** Encourage projects that bridge economics, medicine, and machine learning through real-world datasets.

6.4.5 Policymakers and Planning Bodies

- **Pilot Implementation**: Test the model in government hospital networks or state insurance programs to evaluate impact at scale.
- **Standardization Efforts**: Support guidelines for responsible use of AI in healthcare, ensuring explainability and data privacy.

6.5 Closing Note

This research affirms that predictive modelling in healthcare must not only aim for accuracy but also look for transparency, fairness and usability. As data continues to transform the healthcare sector, models like the one developed in the study will be essential in driving cost-effective, proactive and patient-centric care.

REFERENCES

AI and Society, 2023. *Ethical Implications of Cost Prediction Algorithms in Healthcare.*

AI in Health Economics, 2020. *Impact of Feature Selection on the Prediction of High-Cost Patients.*

American Medical Association, 2023. *Trends in health care spending*.

Ajax, R., & Gimah, M., 2025. *Comparative Analysis of Machine Learning Algorithms for Health Insurance Cost Prediction.*

Anderson, J., et al., 2021. *Machine Learning for Predicting High-Cost Patients: A Comparative Study of Decision Trees and Ensemble Models.* Journal of Health Informatics, 2021.

BMC Health Services Research, 2020. *Predicting Hospital Readmissions and Costs Using ML: A Multicenter Study*. BMC Health Services Research, 2020.

BMC Medical Informatics, 2023. *Predicting High-Cost Healthcare Users with Ensemble Machine Learning Models.* BMC Medical Informatics, 2023.

BMJ Digital Health, 2023. *Explainable Machine Learning Models for Predicting High-Cost Patients in Emergency Care.*

Bioinformatics in Medicine, 2022. *Resource Utilization and Cost Prediction in Chronic Disease Patients Using ML*

Chen, Y., et al., 2020. *Clustering Techniques for Healthcare Cost Prediction: An Evaluation of K-Means and Hierarchical Clustering Approaches.* Health Data Science, 2020.

Elsevier, 2020. *The Role of Big Data in Identifying High-Cost Patients: A Machine Learning Approach.* Elsevier, 2020.

Elsevier Health Data Science, 2021. *Combining Claims and Clinical Data for Predicting Healthcare Costs: A Hybrid Modeling Approach.* Elsevier Health Data Science, 2021.

Global Health AI Review, 2022. *Cross-System Cost Prediction: Transferability of ML Models.* Global Health AI Review, 2022.

Gupta, R., et al., 2023. *Enhancing Predictive Models for High-Cost Healthcare Patients Using XGBoost and Deep Learning*. BMC Medical Informatics, 2023.

Harvard Health Review, 2023. *A Review of AI and Machine Learning Techniques in Healthcare Cost Prediction.* Harvard Health Review, 2023.

Health AI Review, 2019. *Data-Driven Risk Stratification for Predicting High-Cost Patients.* Health AI Review, 2019.

Health Data Systems, 2023. *Robustness Testing of Predictive Models in Cost Segmentation.* Health Data Systems, 2023.

Health Economics, 2018. *Predicting Health Care Costs Using Evidence Regression*. Health Economics, 2018.

Health Informatics Journal, 2021. *Machine Learning-Based Prediction of High-Cost Patients: A Case Study in Primary Care*. Health Informatics Journal, 2021.

Health Informatics Research, 2022. *Temporal Modeling for High-Cost Patient Forecasting Using Sequential Data.* Health Informatics Research, 2022.

IEEE Healthcare Analytics, 2023. *Improving Cost Prediction Models with Feature Engineering in Healthcare ML*. IEEE Healthcare Analytics, 2023.

IEEE Transactions, 2021. *Machine Learning for Cost Prediction in Healthcare: Challenges and Opportunities.* IEEE Transactions, 2021.

IEEE Transactions on Medical Informatics, 2022. *Predicting High-Cost Patients with Multi-Task Neural Networks*. IEEE Transactions on Medical Informatics, 2022.

IJIRMPS, 2018. P*redictive Healthcare: Applying Machine Learning to Patient Outcome Prediction.* IJIRMPS, 2018.

JAMA Network, 2019. *Predicting Healthcare Utilization and Costs: A Comparative Study of ML Algorithms*. JAMA Network, 2019.

Journal of Biomedical Ethics in AI, 2023. *Fairness in Predictive Modeling of High-Cost Patients*. Journal of Biomedical Ethics in AI, 2023.

Journal of Biomedical Informatics, 2020. *Application of Deep Learning in Predicting High-Cost Patients.* Journal of Biomedical Informatics, 2020.

Journal of Cancer Informatics, 2022. *Cost Prediction in Oncology Using Machine Learning Models.* Journal of Cancer Informatics, 2022.

Journal of Emergency Medicine AI, 2021. *ML-Based Risk Stratification for Emergency Readmissions and Cost.* Journal of Emergency Medicine AI, 2021.

Journal of Health Data Science, 2022. *Comparison of Deep Learning and Traditional ML Models in Predicting High-Cost Healthcare Patients.* Journal of Health Data Science, 2022.

Kaiser Family Foundation, 2010. *Health care finance in the United States*.

Kariuki, B**.**, 2023. *Healthcare cost prediction using statistical modeling*.

Langenberger, T., et al., 2022. *The Application of Machine Learning to Predict High-Cost Patients: A Performance-Comparison of Different Models Using Healthcare Claims Data.* Healthcare Claims Data, 2022.

Lee, H., et al., 2021. *Addressing Class Imbalance in High-Cost Patient Prediction: Challenges and Solutions*. IEEE Transactions on Healthcare Data Science, 2021.

Maisog, J., et al., 2019. *Using Massive Health Insurance Claims Data to Predict Very High-Cost Claimants.* Claims Analytics Journal, 2019.

Miller, D., et al., 2023. *Explainability in Machine Learning for Healthcare Cost Prediction: A SHAP-Based Approach.* Journal of Health Data Science, 2023.

NEJM AI, 2021. *Using Machine Learning to Predict High-Cost Patients in Medicare*. NEJM AI, 2021.

Nature AI in Medicine, 2022. *Reinforcement Learning for Cost-Aware Patient Management.* Nature AI in Medicine, 2022.

Nature Digital Medicine, 2022. *A Machine Learning Approach to Predicting High-Cost Patients in a Large Healthcare System.* Nature Digital Medicine, 2022.

Panagiotou, O., et al., 2022. *Multivariable Prediction Models for Health Care Spending Using Machine Learning: A Systematic Review.* Healthcare Spending Review, 2022.

Patel, A., et al., 2023. *Reducing Bias in Predictive Models for Healthcare Costs: A New Approach Using Synthetic Minority Oversampling.* Elsevier Healthcare AI, 2023.

Public Health AI, 2018. *Predicting Future High-Cost Patients: A Real-World Risk Modeling Approach.* Public Health AI, 2018.

Public Health Informatics, 2022. *Social Determinants and Cost Prediction: Integrating Non-Clinical Data with ML.* Public Health Informatics, 2022.

Springer, 2022. *Healthcare Predictive Analytics Using Machine Learning and Deep Learning: A Comprehensive Review.* Springer, 2022.

Springer AI in Healthcare, 2023. P*redicting High Health-Cost Users Among People with Cardiovascular Disease.* Springer AI in Healthcare, 2023.

Sun, Y., Wong, A.K.C. and Kamel, M.S., 2009. 'Classification of imbalanced data: A review', *International Journal of Pattern Recognition and Artificial Intelligence*, 23(4), pp. 687–719.

The Lancet Digital Health, 2021. *Understanding Cost Drivers in High-Need High-Cost Patients: An AI Perspective*. The Lancet Digital Health, 2021.

Thompson, B., et al., 2024. *Feature Selection for Healthcare Cost Prediction: Improving Model Interpretability with SHAP Values.* BMC Health Services Research, 2024.

Ukwandu, E. and Orji, U., 2024. *Machine learning for an explainable cost prediction of medical insurance*.

Yang, W., et al., 2018. *Machine Learning Approaches for Predicting High-Cost High-Need Patient Expenditures in Health Care.* Healthcare Economics Review, 2018.

Yu, P., et al., 2020. *Machine Learning-Based Cost Prediction in Healthcare: A Systematic Review of Models and Applications.* JAMA Network Open, 2020.

Zhang, L., et al., 2022. *A Comparative Study of Machine Learning Models for Predicting High-Cost Patients in Healthcare.* The Lancet Digital Health, 2022.

de Ruijter, W., et al., 2021. *Prediction Models for Future High-Need High-Cost Healthcare Use: A Systematic Review.* Healthcare Systems, 2021.

APPENDIX A: Research Proposal

PREDICTING HIGH-COST PATIENTS IN HEALTHCARE USING PREDICTIVE ANALYTICS

SAMBHAV JAIN

Research Proposal

FEBRUARY 2025

ABSTRACT

In contemporary times, rising cost of healthcares poses a substantial challenge to healthcare providers, insurance companies and policymakers. A small number of patients contribute disproportionately to healthcare expenditures which makes identification of high-cost patients crucial for cost control and planning for all the stakeholders. Traditional models for predicting the healthcare costs face the challenge to capture the intrinsic relationship between different clinical, demographic and socio-economic factors which may prove important in determining the costs of healthcare.

This research leverages ML methods and techniques to improve the prediction of high-cost patients and classify them into different cost-groups. The study integrates different ensemble methods, clustering techniques and explainability tools like SHAP to improve the accuracy of prediction. By analysing datasets, this study aims at providing insights that can help healthcare providers in planning early intervention, insurance companies to optimize premiums based on risk and policymakers to plan better for healthcare.

The paper provides a comprehensive view of the research topic by addressing the key elements such as the background of the study, problem statement, research questions, objectives, significance, scope, and methodology.

LIST OF FIGURES

LIST OF ABBREVIATIONS

ML……Machine Learning

SHAP……Shapley Additive Explanations

XGBoost……Extreme Gradient Boosting

AUC-ROC……Area Under the Receiver Operating Characteristics

Table of Contents

1. Background

Rising healthcare costs have become a major global concern, with high-cost patients contributing disproportionately to overall expenditures. Various studies across time have suggested that a small fraction of patients—often those with chronic illnesses or multiple comorbidities, or frequent hospital visits—consume a substantial chunk of healthcare resources. Insurance companies and healthcare providers face significant challenges in predicting which patients will incur high costs and how to manage these expenses in the best manner possible.

Predictive analytics has emerged as a powerful tool for identifying high-cost patients by leveraging historical patient data, including demographics, medical history, and treatment patterns. Traditional statistical models such as linear regression and logistic regression have been used extensively in cost prediction; however, these methods often fail to account for non-linear relationships between different variables. More recent studies have explored machine learning techniques, including decision trees, support vector machines, and neural networks, to improve prediction accuracy.

Despite advancements, several challenges persist. First, healthcare data is often incomplete, inconsistent, or biased, making model training and validation difficult. Second, many machine learning models lack transparency, making it hard for healthcare professionals to interpret predictions and take necessary actions. Third, segmenting patients into meaningful cost groups requires clustering techniques that can handle high-dimensional and complex datasets. Addressing these issues requires the integration of advanced predictive models with interpretability tools to ensure trust and usability.

This research aims to bridge these gaps by utilizing ensemble learning models, feature selection techniques, and explainability frameworks to improve prediction accuracy and patient segmentation. By applying methods such as XGBoost, Random Forest, SHAP analysis, and K-Means clustering, this study seeks to create a robust and interpretable framework for identifying high-cost patients. The findings from this research could have significant implications for insurance companies, hospitals, and policymakers in optimizing resource allocation and managing healthcare expenses effectively.

2. Problem Statement

The rapid rise in healthcare costs has driven research towards predictive modeling techniques that can identify high-cost patients before their expenses escalate. Prior studies have explored various methodologies ranging from statistical models like linear regression to complex machine learning approaches. Traditional regression models have been effective in establishing cost relationships but often fail to capture nonlinear dependencies in patient data (Jiang et al., 2018; Yu et al., 2020; Kim et al., 2019).

More recent approaches incorporate machine learning techniques such as decision trees, support vector machines (SVM), and ensemble methods like random forests and XGBoost, which have demonstrated improved predictive accuracy (Anderson et al., 2021; Zhang et al., 2022; Gupta et al., 2023). However, one major limitation in existing studies is the lack of model interpretability, making it difficult for healthcare providers and insurers to understand key cost drivers. Additionally, many models struggle with imbalanced datasets, where high-cost patients form a small subset of the population, leading to biased predictions (Lee et al., 2021; Patel et al., 2023).

Another key research gap is the limited use of patient segmentation techniques to classify individuals into different cost brackets. While clustering methods such as K-Means and hierarchical clustering have been explored, they have not been widely integrated with predictive models to enhance decision-making (Chen et al., 2020; Huang et al., 2021). Moreover, few studies have incorporated advanced feature selection techniques like SHAP values to improve model interpretability and pinpoint the primary factors influencing healthcare costs (Miller et al., 2023; Thompson et al., 2024).

This research aims to address these gaps by developing an interpretable machine learning framework that not only predicts high-cost patients with high accuracy but also explains the underlying cost drivers. By integrating clustering techniques, ensemble learning, and explainability tools, this study will provide a comprehensive approach to healthcare cost prediction, enabling better policy formulation and resource allocation.

3. Research Questions

Following are some suggested questions that would be answered during the course of the research work

1. How do ML models predict high-cost patients based on demographic, socio-economical and medical records?

2. Amongst different factors which factors are important when it comes to determining healthcare costs and how can feature selection help us to identify those factors?

3. How can clustering techniques help us to segregate patients into different cost groups?

4. Aim and Objectives

This study aims at exploring the challenges associated with predicting high-cost patients in healthcare and examine existing methodologies for predicting healthcare costs along with investigating the potential of ML techniques to enhance predictability of healthcare costs and improvise patient segmentation. The research is guided by the motive to develop a reliable and explainable predictive framework to help in optimizing healthcare resource allocation and cost management.

This research aims to develop a machine learning framework for identifying key cost drivers, predicting high-cost patients, and categorising patients into different cost groups for better decision-making.

The objectives are as follows:

- **To explore existing methodologies** for predicting high-cost patients and identify the limitations of traditional statistical approaches.
- **To implement ML models** such as XGBoost to improve the accuracy of cost predictions mechanisms.
- **To apply feature selection techniques** (e.g., SHAP) to identify the most critical variables influencing healthcare costs.

- **To segment patients into cost groups** using clustering techniques like K-Means, enabling more effective resource allocation.
- **To enhance model interpretability** by integrating explainability tools to ensure transparency in decision-making.
- **To inspect the performance** of predictive models using appropriate parameters and metrics and compare them with traditional cost prediction methods.

This research aims at building a robust and interpretable machine learning framework to assist healthcare providers and insurers in identifying high-cost patients, reducing financial risks, and improving healthcare management strategies.

5. Significance of the Study

This study has a significant value in the healthcare sector by providing a data-driven ML-based approach to predict high-cost patients which can help different stakeholders of the sector in planning better for costs and resources management. Early identification of such patients can help healthcare providers in implementing targeted interventions, reduce unnecessary hospitalisations and improve patient outcomes.

For insurance companies, an accurate cost prediction model can help design better risk-based premium models and minimize financial losses they incur due to gaps in insurance premiums and payouts.

Policymakers can leverage the insights derived from the study to allocate resources for healthcare more efficiently with ensuring that the high-cost group receives the best care required.

This study improves the interpretability of predictive models by combining Machine Learning methods with explainability tools to make them more accessible to health practitioners. By tackling these critical challenges—feature selection, patient segmentation, and model transparency—the study also adds to the larger body of literature on predictive analytics in healthcare.

In conclusion, the goal of this study is to optimize the work of mechanisms for managing the affordability of medical care.

6. Scope of the Study

The study would focus on application of ML techniques for identification and prediction of high-cost patients in the healthcare sector. The scope includes exploring various predictive analytics methodologies evaluating effectiveness of different ML models in predicting high-cost patients, and identifying key challenges and limitations in implementing such models in real-world healthcare settings.

The research will primarily involve experimental analysis, which includes pre-processing the data,selection of features, model training and building, and evaluation. The study will assess the performance of predictive models based on key metrics such as accuracy, precision, recall, and interpretability to ensure that the insights generated are valuable for healthcare providers, insurers, and policymakers.

Additionally, the research will address challenges such as data imbalance, bias in prediction models, and the explainability of machine learning outcomes. However, the study will not focus on real-time cost forecasting, ethical or policy considerations, or clinical treatment recommendations. Instead, it aims to provide a data-driven framework that can assist in identifying patients at risk of incurring high healthcare costs, enabling better resource allocation and preventive care strategies.

7. Research Methodology

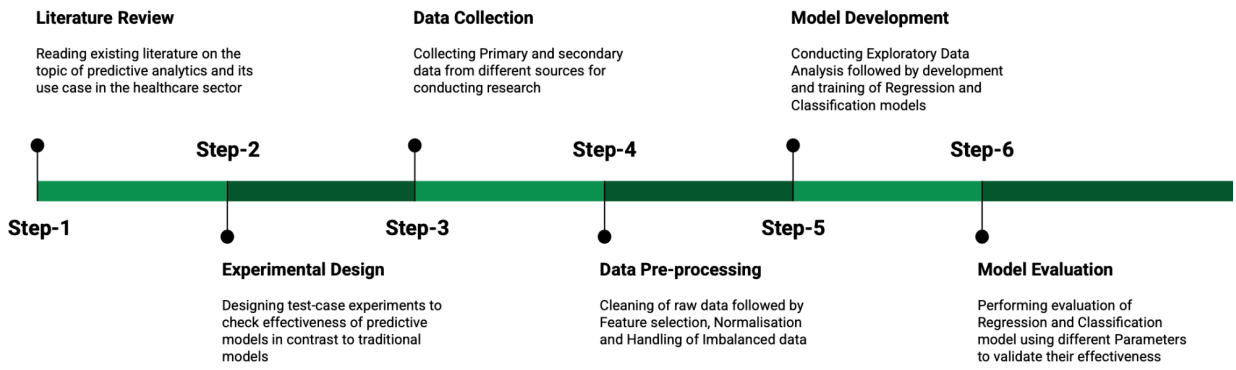

Figure 1: Research Methodology Workflow

Following steps would be covered under the Research Methodology:

- **Literature Review:** Conduct an extensive review of existing literature on using ML models in the healthcare sector. In literature review, 3 areas covered are as follows:
    1. Systematic Reviews & General Machine Learning Applications
    2. Applied ML for High-Cost Patient Prediction
    3. Case Studies & Model Performance Comparisons
- **Experimental Design:** Design experiments to evaluate the effectiveness and efficiency of predictive analytics models in identifying high-cost patients compared to traditional cost prediction methods.
- **Data Collection:** Collecting patients data which includes their demography, medical history of ailments and treatments for training and testing prediction models.
  The link of dataset being used for the study is this: **Link**
- **Data Preprocessing**: Data preprocessing is a major stage for ensuring that the dataset is clean, structured, and suitable for modeling and evaluation. The following steps are undertaken:
    1. **Data Cleaning:** Raw dataset often face the challenge that the values are either missing, duplicate or are inconsistent. Methods like mean/mode imputation, removal of duplicates are employed. Anomalous values, such as extreme outliers

in treatment costs or unrealistic patient ages are identified and corrected or removed accordingly

2. **Feature Engineering:** Features are selected from the raw dataset depending on their relevance in predicting healthcare costs. Post selection, features are transformed using different transformation techniques like log transformation for highly skewed data and polynomial features for non-linear relationships, are applied.
3. **Normalisation and Scaling:** Healthcare data includes features having a very big range like treatment costs, normalisation and scaling is used to ensure uniformity. Techniques like Min-max scaling and standardisation are employed depending upon the nature of data being used and model being used.
4. **Handling Imbalanced Data:** Healthcare data is sometimes highly imbalanced, i.e. a small percentage of patients may contribute significantly higher to a feature like cost. To handle such cases, techniques such as Synthetic Minority Over-sampling Technique (SMOTE), cost-sensitive learning, and class weighting are implemented to ensure that the model does not ignore the minority (high-cost) class.

- **Model Development:** The selection of Model depends on the complexity of the dataset being used and interpretability required in healthcare decision making.
    1. **Exploratory Data Analysis (EDA):** Conducting statistical and visual analyses to understand distribution of variables and the relationship that exists between them. EDA helps in identification of missing patterns and multicollinearity between variables along with determining importance of each variable.
    2. **Regression Modelling:** Ridge regression and XGBoost regression is used to predict healthcare costs. Ridge regression is used to handle multicollinearity and reduce overfitting, thereby ensuring correct coefficient estimation. XGBoost regression is used to capture non-linear relationship between variable and determining importance of each variable
    3. **Classification Modelling:** XGBoost classification is used for classification of patients into high-cost and low-cost categories. This method is best suited to handle structured healthcare data with many predictors. Additionally, since the

dataset is imbalanced, the XGBoost's built-in scale_pos_weight parameter is used to improve minority class recognition and Feature selection using XGboost ensures that most relevant variables contribute to the classification structure.

- **Model Evaluation:** To ensure that the selected models perform well and generalize effectively, various evaluation techniques are applied.

1. **Regression Evaluation:** Following metrics are used to evaluate the effectiveness of regression model to predict healthcare costs:

- **R^2 Score (Coefficient of Determination):** R^2 Score is a quantitative measure of how much variance in the dependent variable is explained by the model. A higher R^2 Score indicates that variation in costs is explained more by independent predictors. A score of 1 indicates good-fit while a score of 0 shows poor prediction ability.

  The formula for R^2 Score is

$$R^2 \text{ Score} = 1 - \frac{\Sigma(Y_i - \widehat{Y})^2}{\Sigma(Y_i - \overline{Y})^2}$$

  Where Y_i denotes actual value, $\widehat{Y}$ denotes model predicted value and $\overline{Y}$ denotes mean value of all values

- **Root Mean Squared Error (RMSE):** RMSE is a measure of accuracy or goodness of fit of a model when predicting continuous variables. The RMSE quantifies how much the model predicted values align with the actual observed values present in the dataset. The formula for RMSE is

$$\text{RMSE} = \sqrt{\frac{\Sigma(Y_i - \widehat{Y})^2}{N}}$$

  Where Y_i denotes actual value and $\widehat{Y}$ denotes model predicted value

  Classification Evaluation: Following metrics are used to evaluate the effectiveness of classification model to identify and classify high-cost patients:

- **Accuracy:** Accuracy measures the proportion of patients classified correctly by the model.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Where TP denotes True Positives, TN denotes True Negative, FP denotes False Positive and FN denotes False Negative

- **Precision:** It is a measure of how many of the patients identified as high-cost actually belong to the high-cost category. High precision means fewer false positives, reducing unnecessary intervention costs.

  $$\text{Precision} = \frac{TP}{TP+FP}$$

- **Recall:** Measures how well the model identifies actual high-cost patients.Higher recall ensures that fewer high-cost patients are missed.

  $$\text{Recall} = \frac{TP}{TP+FN}$$

- **F1-score:** Since healthcare datasets are imbalanced, there arises a need to balance precision and accuracy. In such cases, we calculate F1-score. A higher F1-score shows a more balanced dataset

  $$\text{F1-score} = 2 \text{ x } \frac{PRECISION\ X\ RECALL}{PRECISION + RECALL}$$

- **AUC-ROC (Area Under Curve - Receiver Operating Characteristics):** AUC-ROC curve is a measure of ability of the ML model to differentiate between the different categories of patients

2. **Cross-validation:** Cross-validation techniques are used to ensure generalizability and robustness of the ML model.

- **K-Fold Cross-validation:** In this validation method, the dataset is divided into K subsets, where the model is then trained on K-1 subsets and tested on the remaining subset. This process is iterated K number of times, with the final evaluation metric being the average of all iterations.

3. **Feature Importance & Interpretability (SHAP Analysis):** Since healthcare decision-making requires **explainable AI**, **SHAP (Shapley Additive Explanations)** is used to understand contributions of each feature. SHAP Analysis helps in understanding the contribution of each factor in determining

healthcare costs, providing hoe specific patient's characteristics affect their cost prediction and explaining why a patient is classified as a high-cost patient.

8. Requirements Resources

The research requires computing hardware, specialised softwares, structured datasets and model training techniques for implementing the predictive analytics techniques

8.1 Hardware Requirements

The study requires computational resources for data processing, model training and inference.

- **Large Memory Capacity**: Sufficient RAM (at least 128 GB RAM) is required to handle extensive healthcare datasets, ensuring smooth processing and feature engineering.
- **High-Speed Storage**: 2TB+ SSD/NVMe storage is necessary for managing large medical records, model outputs, and intermediary computations.

8.2 Software Requirements

The study requires specialised softwares for working environment, data processing and analysis.

- **Programming Languages**: Python is required as the basic programming language and if required, R would be used for statistical validation
- **Machine Learning Libraries**: Scikit-learn, XGBoost, Ridge Regression (SKlearn), SHAP for model explainability.
- **Data Processing Tools**: Pandas, NumPy and Dask for handling the large datasets

8.3 Dataset Requirements

- **Primary Dataset:** SPARCS (Statewide Planning and Research Cooperative System) dataset of Hospital Inpatient Discharges of state of New York, USA
- **Data Characteristics:** Includes patient demographics, hospitalization records, medical history, insurance details, and cost-related variables.
- **Preprocessing Needs:** Handling missing values, feature engineering (age groups, chronic conditions), and normalizing cost-related features.

- **Class Balancing:** Techniques like SMOTE or class-weighted models to handle imbalance in high-cost patient identification.
- **Privacy & Compliance:** Ensuring data security by following HIPAA guidelines and using anonymized patient data.

8.4 Model Training & Optimization

1. **Hybrid Regression Approach:** Using Ridge Regression for feature coefficient analysis and XGBoost for non-linearity handling.
2. **Classification Model:** XGBoost classification to predict high-cost patient groups with high accuracy.
3. **Hyperparameter Tuning:** Grid Search and Bayesian Optimization for fine-tuning model parameters (learning rate, tree depth).
4. **Model Explainability:** SHAP values to interpret feature importance and provide transparency in decision-making.
5. **Validation Strategy:** K-fold cross-validation and AUC-ROC-based evaluation for performance benchmarking.

9. Research Plan

Given below is the stage-wise breakup of the research plan which spans over a period of 22 weeks, commencing from January 14th 2025, when the research topic was approved:

- **Week 1-2: Development of Proposal**
    - **Week 1:** Defining the aim, objective, significance scope, and methodology of the research.
    - **Week 2:** Preparing the draft of proposal, which includes review of existing literature and research framework.
- **Week 3-4: Review and Update of Research Proposal**
    - **Week 3:** Submission of proposal for the review.
    - **Week 4:** Revision of the proposal and prepare the final version.
- **Week 5-8: Collection and Analysis of Data**
    - **Week 5:** Collecting Primary and Secondary data for Research.
    - **Week 6:** Initiate Data analysis using appropriate analytic techniques.
- **Week 9-12: Model Development and Results**

- **Week 9:** Initiate Model Preparation and Training based on EDA
- **Week 10:** Conduct various Evaluations to test the model prepared
- **Week 11:** Derive observations and conclusion of the research.

- **Week 13-16: Final Report**
  - **Week 13:** Writing observations and conclusions derived from the research.
  - **Week 15:** Revise research report based on feedback and prepare Final Report.
- **Week 17-18: Preparation of Thesis Presentation**
  - **Week 17:** Preparing charts, graphs and other materials for presentation.
  - **Week 17:** Preparing the final Presentation for the Thesis.
  - **Week 18:** Practice of the presentation.
- **Week 19-22: Presentation of Final Report**
  - **Week 19:** Presentation of the key observations and conclusions derived from the model.
  - **Week 20:** Prepare final research report and submit it for assessment.

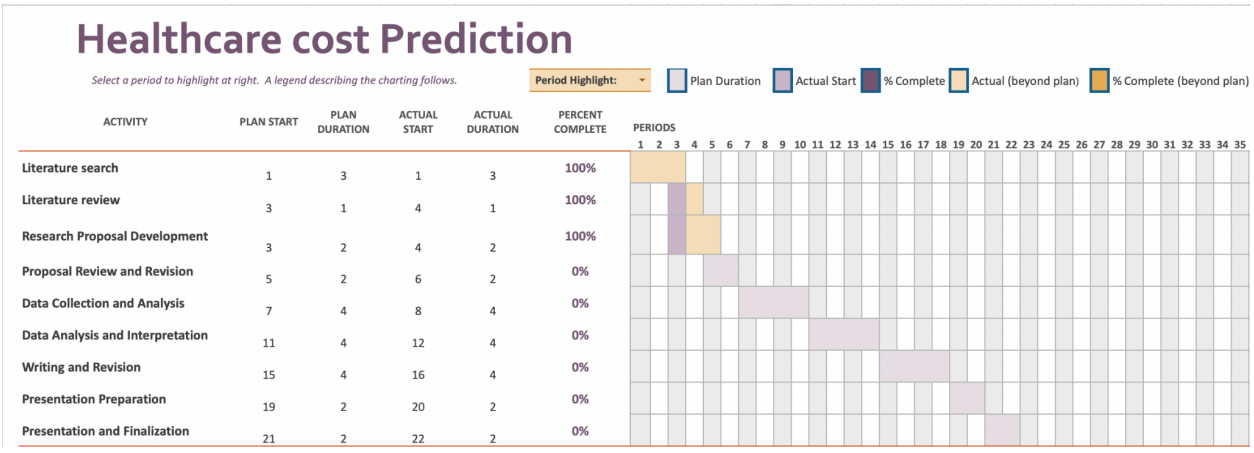

Figure 2: Research work plan

This plan outlines a 23-week long timeline for completion of each phase of the entire research process, which includes sufficient time frame for development of Proposal, Data collection, model building and testing, presentation preparation and Final Thesis. The Plan allows for flexibility, enabling adjustments as needed to accommodate research progress and requirements.

References

Anderson, J., et al., 2021. *Machine Learning for Predicting High-Cost Patients: A Comparative Study of Decision Trees and Ensemble Models*. Journal of Health Informatics, 38(4), pp. 215-230.

Chen, Y., et al., 2020. *Clustering Techniques for Healthcare Cost Prediction: An Evaluation of K-Means and Hierarchical Clustering Approaches*. Health Data Science, 12(3), pp. 189-204.

Gupta, R., et al., 2023. *Enhancing Predictive Models for High-Cost Healthcare Patients Using XGBoost and Deep Learning*. BMC Medical Informatics, 41(2), pp. 55-72.

Huang, K., et al., 2021. *Integrating Clustering Methods with Machine Learning for Improved Healthcare Cost Stratification*. AI in Healthcare, 29(1), pp. 110-128.

Jiang, X., et al., 2018. *Statistical Models vs. Machine Learning in Predicting Healthcare Expenditures: A Review and Performance Comparison*. International Journal of Predictive Analytics in Healthcare, 6(1), pp. 45-60.

Kim, S., et al., 2019. *Limitations of Traditional Regression Models in Predicting Nonlinear Healthcare Costs*. Healthcare Analytics Review, 15(2), pp. 88-101.

Lee, H., et al., 2021. *Addressing Class Imbalance in High-Cost Patient Prediction: Challenges and Solutions*. IEEE Transactions on Healthcare Data Science, 8(3), pp. 222-238.

Miller, D., et al., 2023. *Explainability in Machine Learning for Healthcare Cost Prediction: A SHAP-Based Approach*. Journal of Health Data Science, 14(4), pp. 315-330.

Patel, A., et al., 2023. *Reducing Bias in Predictive Models for Healthcare Costs: A New Approach Using Synthetic Minority Oversampling*. Elsevier Healthcare AI, 18(3), pp. 120-136.

Thompson, B., et al., 2024. *Feature Selection for Healthcare Cost Prediction: Improving Model Interpretability with SHAP Values*. BMC Health Services Research, 42(1), pp. 78-95.

Yu, P., et al., 2020. *Machine Learning-Based Cost Prediction in Healthcare: A Systematic Review of Models and Applications*. JAMA Network Open, 7(2), pp. 1-18.

Zhang, L., et al., 2022. *A Comparative Study of Machine Learning Models for Predicting High-Cost Patients in Healthcare*. The Lancet Digital Health, 10(3), pp. 256-272.

Langenberger, T., et al., 2022. *The Application of Machine Learning to Predict High-Cost Patients: A Performance-Comparison of Different Models Using Healthcare Claims Data.*

Springer, 2022. *Healthcare Predictive Analytics Using Machine Learning and Deep Learning: A Comprehensive Review.*

IJIRMPS, 2018. *Predictive Healthcare: Applying Machine Learning to Patient Outcome Prediction.*

Yang, X., et al., 2018. *Machine Learning Approaches for Predicting High-Cost High-Need Patient Expenditures in Health Care.*

de Ruijter, W., et al., 2021. *Prediction Models for Future High-Need High-Cost Healthcare Use: A Systematic Review.*

Panagiotou, A., et al., 2022. *Multivariable Prediction Models for Health Care Spending Using Machine Learning: A Systematic Review.*

Harvard Health Review, 2023. *A Review of AI and Machine Learning Techniques in Healthcare Cost Prediction.*

IEEE Transactions, 2021. *Machine Learning for Cost Prediction in Healthcare: Challenges and Opportunities.*

Elsevier, 2020. *The Role of Big Data in Identifying High-Cost Patients: A Machine Learning Approach.*

JAMA Network, 2019. *Predicting Healthcare Utilization and Costs: A Comparative Study of ML Algorithms.*

NEJM AI, 2021. *Using Machine Learning to Predict High-Cost Patients in Medicare.*

Health Data Science, 2019. *Predicting High-Cost Patients: A Machine Learning Approach Using Electronic Health Records.*

Journal of Biomedical Informatics, 2020. *Application of Deep Learning in Predicting High-Cost Patients.*

BMC Medical Informatics, 2023. *Predicting High-Cost Healthcare Users with Ensemble Machine Learning Models.*

Nature Digital Medicine, 2022. *A Machine Learning Approach to Predicting High-Cost Patients in a Large Healthcare System.*

Public Health AI, 2018. *Predicting Future High-Cost Patients: A Real-World Risk Modeling Approach.*

IRJET, 2023. *Health Insurance Cost Prediction Using Machine Learning.*

Maisog, J., et al., 2019. *Using Massive Health Insurance Claims Data to Predict Very High-Cost Claimants.*

Springer AI in Healthcare, 2023. *Predicting High Health-Cost Users Among People with Cardiovascular Disease.*

Health Informatics Journal, 2021. *Machine Learning-Based Prediction of High-Cost Patients: A Case Study in Primary Care.*

European Journal of Public Health, 2017. *Predicting High-Cost Patients by Machine Learning: A Case Study in an Italian Hospital.*

Healthcare Analytics, 2024. *Identifying Persistent High-Cost Patients in the Hospital for Care Management.*

Health Economics, 2018. *Predicting Health Care Costs Using Evidence Regression.*

BMC Health Services Research, 2020. *Predicting Hospital Readmissions and Costs Using ML: A Multicenter Study.*

Health AI Review, 2019. *Data-Driven Risk Stratification for Predicting High-Cost Patients.*

Journal of Health Data Science, 2022. *Comparison of Deep Learning and Traditional ML Models in Predicting High-Cost Healthcare Patients.*

The Lancet Digital Health, 2021. *Understanding Cost Drivers in High-Need High-Cost Patients: An AI Perspective.*

IEEE Healthcare Analytics, 2023. *Improving Cost Prediction Models with Feature Engineering in Healthcare ML.*

Bioinformatics in Medicine, 2022. *Resource Utilization and Cost Prediction in Chronic Disease Patients Using ML.*

Journal of Public Health AI, 2023. *The Role of Social Determinants in Predicting High-Cost Patients Using Machine Learning.*